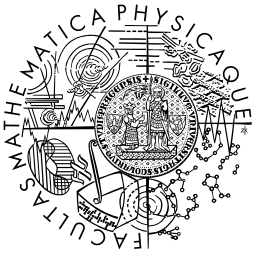


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Anna Dvořáková

Mapování gregoriánského repertoáru

Ústav formální a aplikované lingvistiky

Vedoucí bakalářské práce: MgA. Jan Hajič, Ph.D.
Studijní program: Informatika

Praha 2024

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

Mé nejhlubší poděkování na tomto místě zcela jistě náleží vedoucímu mé práce, MgA. Janu Hajičovi, Ph.D., za pomoc s vytyčením jinak klikatého záhonu mé bakalářské práce a dále s jeho trpělivou údržbou a úpravou, bez něj by se žádné pokusy o pěstování mrkve nekonaly.

Dále nemohu opomenou prof. Davida Ebena z ÚHV FF UK, jehož trpělivost s vysvětlováním humanitní látky, byť poučenému, matfyzákovi, mi tolik pomohla a snad přinesla alespoň pár růžových květů.

V neposlední řadě pak patří velký dík rodině a přátelům, kteří v prošlém roce prokázali velkou shovívavost k mým omezeným časovým možnostem, snášeli mé vykládání o trochu prapodivné bakalářce a říkali: „Jsi šikovný ježeček!“ vždy, když jsem to potřebovala slyšet. Zasloužíte si kytičku. Kač.

Název práce: Mapování gregoriánského repertoáru

Autor: Anna Dvořáková

Ústav: Ústav formální a aplikované lingvistiky

Vedoucí bakalářské práce: MgA. Jan Hajič, Ph.D., Ústav formální a aplikované lingvistiky

Abstrakt: Gregoriánský chorál, zpívaný jednohlas a součást jádra latinské liturgie, je základní částí evropské hudební historie, důležitou a dobře zachovanou součástí středověkého kulturního dědictví a také populárním předmětem studia muzikologů. Jedním z hlavních problémů zkoumaných v gregorianistice je variabilita chorálního repertoáru v pramenech navzdory vysoké standardizaci napříč celou latinsky hovořící Evropou. Muzikologové a historici zkoumají existenci, rozsah a vývoj určitých subtradic, které ukazují, jak ve středověku probíhalo šíření kulturních novot. Díky síti databází Cantus Index je k výpočetnímu výzkumu potenciálních tradic zpřístupněno více než 800,000 digitálních katalogových záznamů. V této práci k pokusům o rozpoznání repertoárových tradic používáme metody shlukové analýzy, detekce komunit a modelování témat. Otázka existence tradic se ukázala být potenciálně problematickou a vyžaduje další evaluaci příslušnými odborníky z řad muzikologů. Za tímto účelem práce přináší softwarový nástroj, který umožňuje geografickou vizualizaci výsledků hledání tradic.

Klíčová slova: gregoriánský chorál, digitální muzikologie, geoinformatika

Title: Mapping the Repertoire of Gregorian Chant

Author: Anna Dvořáková

Institute: Institute of Formal and Applied Linguistics

Supervisor: MgA. Jan Hajič, Ph.D., Institute of Formal and Applied Linguistics

Abstract: Gregorian chant, the vocal monody at the core of Latin liturgy, is a fundamental part of European music history, an important part of the medieval cultural heritage with good preservation and a popular filed of study for musicologists. One of the central problems examined in Greogrian chant scholarship is the variability of chant repertoire found in sources despite the high degree of standardization across all of Latin Europe. Musicologists and historians examine the existence, extent, and development of certain sub-traditions that reveal how cultural innovation spread in the Middle Ages. Thanks to the Cantus network of databases, a large amount of more than 800,000 digital catalogue records is available to examine potential traditions computationally. In this work, we use methods of clustering, community detection and topic modelling to try and detect repertoire traditions in this data. The existence of repertoire traditions proves to be potentially problematic and requires further evaluation by relevant musicologists. For this purpose, this thesis also presents software tool that provides geographical visualisation of the results of search for traditions.

Keywords: gregorian chant, digital musicology, geoinformatics

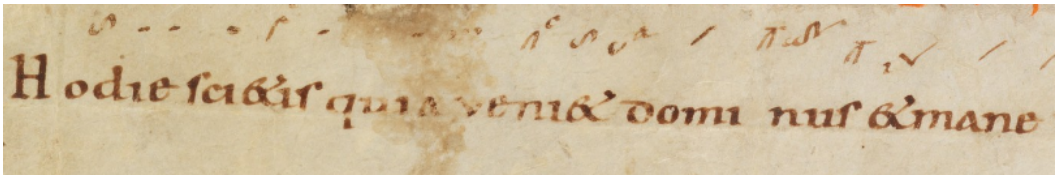
Obsah

Úvod

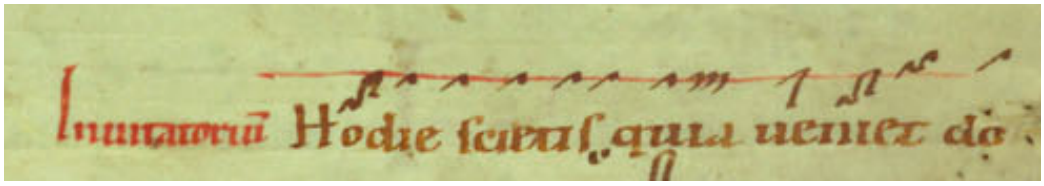
Co mají společného páže servírující oběd knížeti Václavovi (927), dělník na stavbě Karlova mostu (1357) a voják v bitvě na Bílé hoře (1620)? Všichni v kostele slýchali totéž. Hudební složka mše, respektive ještě specifičtěji právě chorál, vypadala pro všechny velice obdobně. Hygienické návyky, poznání vesmíru, móda i jazyk se za tu dobu proměnily, ale základní hudbou v kostele byl stále jednohlasý latinský zpěv nazývaný gregoriánský chorál. Skutečně však slyšeli všichni tři zmiňovaní, obyvatelé středověku a raného novověku, úplně totéž?

Gregoriánský chorál, který je od poč. 9. století univerzálním zpěvem katolické církve, je nejenom jedním ze základních kamenů evropské hudby, ale také kulturním a historickým fenoménem, který hraje důležitou roli na poli dokladů o středověké kultuře. To platí obzvlášť vzhledem k unikátní zachovalosti, dané také institucionálním tlakem na zakonzervování této posvátné tradice v neměnné podobě. Zajímavé je, že i přesto chorál vykazuje zkoumatelné rozdíly.

Studium těchto odlišností je předmětem podoboru gregorianistiky *transmission of chant*, tedy česky přenos či šíření chorálu. Zde lze sledovat dvě roviny – horizontální (skrze prostor, a to nejenom z Říma ven) a vertikální (skrze čas). Jelikož byl chorál po celý raný středověk převážně ústní tradicí a i s výskytem zápisu not stále hrálo významnou roli paměťové uchování – kniha byla jedna, zpěváků, i přes jednohlasou formu, více – nějakou vnitřní variabilitu navzdory oněm konzervačním tlakům očekáváme, podobně jako v libovolné jiné orální tradici; o to těžší je však přenos mapovat. Přesto se muzikologové tážou, jak přenos či šíření probíhaly nebo zdali v rámci repertoáru existují nějaké skupiny z různých míst či časů vykazující společné znaky, které se v jiných oblastech a časech nevyskytují, nějaké skupiny držící tradici.



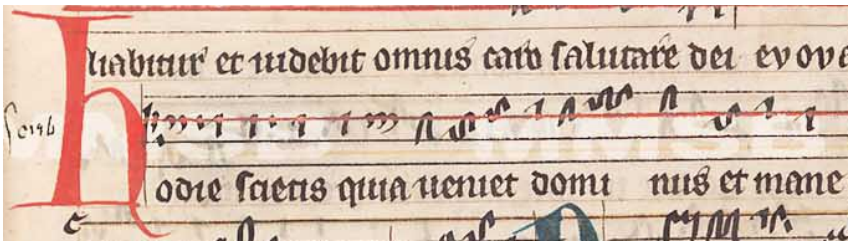
Obrázek 1: Zpěv *Hodie scietis* z CH-SGs 390, f. 041, konec 10. století



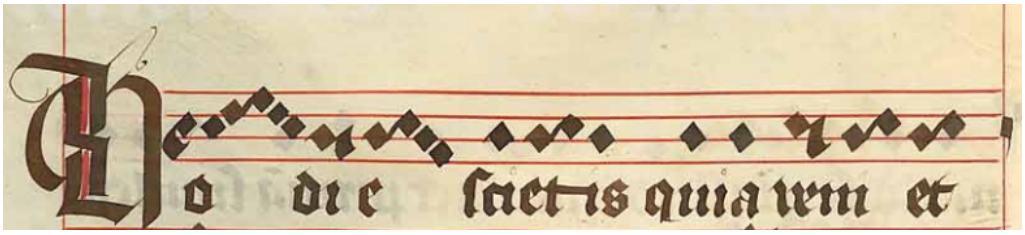
Obrázek 2: Zpěv *Hodie scietis* z A-KN 1010, f. 019r, 12. století

Na obrázcích 1, 2, 3, 4 a 5 vidíme zápisy zpěvu *Hodie scietis* (jedná se o zpěv používaný na Štědrý den) vytvořené v různých dobách, notacích, místech či druzích institucí. Dokonce i melodie se mírně mění, ale text a použití přetrvaly. Do jaké míry je však výskyt *Hodie scietis* v části příslušející oslavě Vánoc v uvedených pramenech reprezentativní pro zbytek Evropy? Nebo se shodují náhodou?

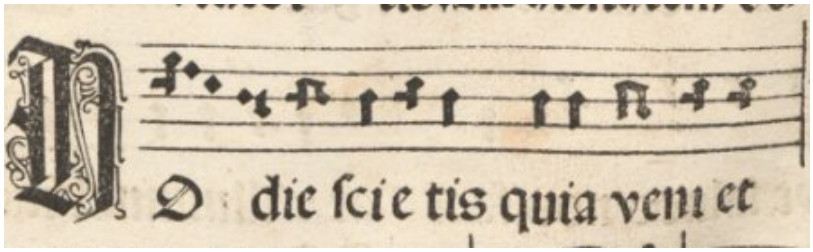
A pokud se v jiných místech a časech používaly pro příslušnou část Vánoc jiné zpěvy, proč tomu tak bylo? Jak se tyto změny v jinak silně zakonzervovaném kulturním fenoménu děly?



Obrázek 3: Zpěv *Hodie scietis* z Cz-Pu XIV B 13, f. 021r, počátek 14. století



Obrázek 4: Zpěv *Hodie scietis* z Cz-Pn XII A 24, f. 027v, počátek 15. století



Obrázek 5: Zpěv *Hodie scietis* z MA Impr. 1537, f. 025r, 16. století

Přenos či šíření chorálu je navíc i přenosem či šířením středověké literární kultury. Spolu s novou částí repertoáru cestoval člověk, který ho uchovával v paměti či později nesl na pergamenu. Lze předpokládat, že cesty minimálně některých takových sledovaly nějakou standardní trasu, a tedy i linii kulturního přejímání, že dotyčný možná nenesl „jen zpěvy“. Výhodou gregoriánského chorálu je jeho vzájemná porovnatelnost napříč velkou geografickou oblastí a časovým obdobím (mj. nepodléhá změnám uměleckých epoch). Poskytuje tak přístupnější možnost hledat cesty přenosu i jiného než pouze chorálního materiálu.

Komunita gregorianistů už mnoho let kompletuje digitální databázi chorálních zpěvů. V rámci ducha *digital humanities* jsme se pro sesbírané netriviální množství tohoto materiálu rozhodli, že by mohlo dávat smysl zaměstnat technologie. Ve výpočetní složce se jedná o informatický výlet za problémy oboru, který, z principu nastíněných výzkumných otázek a dosaženého dosavadního poznání, neumí poskytnou žádné správné řešení, který sám nezná odpovědi a v globálním měřítku mu lze pouhou specifikací otázky pomoci. Práce má navíc za cíl i přínos digitální, a sice v poskytnutí interaktivní geografické vizualizace míst původu

dochovaných pramenů gregoriánského chorálu, která výsledky zaměstnaných informatických přístupů přehazuje zpátky muzikologům, aby využili své oborové znalosti k jejich podrobnému posouzení.

Obrázek 6: Raně středověká ilustrace zobrazující společný přednes

Obrázek 7: Iniciála C vyzdobená zpívajícími mnichy, pozdní středověk

Obrázek 8: Zpívající andělé na renesančním oltářním obrazu v Ghentu (Jan van Eyck)

Hlavní přínosy této práce jsou:

- Provedení prvních pokusů s výpočetním hledáním tradic pomocí tří různých výpočetních přístupů: shlukování, síťová analýza a bayesovský model témat.
- Průzkum limitací existující datové sady dostupné přes síť databází Cantus Index.
- Vývoj nástroje, který zpřístupňuje výpočetní hledání repertoárových tradic muzikologům.

První kapitola práce (1) uvádí základní přehled gregoriánského chorálu do hloubky potřebné pro tuto práci. Následující kapitola (2) přináší přehled souvisejících prací. Ve třetí kapitole (3) představujeme použitou datovou sadu. Čtvrtá kapitola (4) popisuje použité metody výzkumu a představuje provedené experimenty. Následující dvě kapitoly 5 a 6 dokumentují implementovaný nástroj – z uživatelského a z vývojového hlediska.

1. Gregoriánský chorál

Gregoriánský chorál je jednohlasý zpěv používaný k oslavě Boha v západní křesťanské tradici. Jedná se o výraznou součást středověkého kulturního dědictví a zároveň součást dodnes živou. Jakkoliv jde původně o ústní tradici, od devátého století vznikají i textové a notové zápisy, což nám dnes umožňuje snazší analýzu.

Hlavním úkolem církevních jednotek středověku bylo slavení bohoslužeb. Právě během středověku se rozsah liturgie, bohoslužebného programu, notně rozvinul a nabyl na rozsahu i celkové složitosti. Zpívaná forma (často biblickým) textům používaným při bohoslužbě dodávala váhu, oddělovala je od běžného života, vytvářela posvátno[1], podobně jako například prostor kostela se liší od většiny jiných běžných budov a interiérů a ornát kněze je něčím víc než jen obyčejným oděvem. Navíc vzhledem k používání latinských textů mohl být pro většinové evropské obyvatelstvo krásný zpěvný zvuk duchovně vlastně mnohem uchopitelnější než vlastní, přes jazykovou bariéru nedosažitelný, obsah textů.

Tato kapitola vychází z úvodních textů gregorianistiky od Davida Hileyho (Hiley, 2009) a o něco staršího Richarda H. Hoppina (Hoppin, 2007). Dále stojí za zmínku i jim předcházející rozsáhlý výkladový text sepsaný Apelem (Apel, 1958) a také jedno z prvních systematických děl o Gregoriánském chorálu, Wagner (1911).

1.1 Žánry

Jednotlivé zpěvy můžeme rozdělit do svou větších kategorií podle stylu, které dále dělíme na konkrétnější žánry. Jedná se o dělení na liturgický recitativ a volnou kompozici. Toto rozdělení vychází ze snahy najít rovnováhu mezi srozumitelností zpívaného textu a krásou a bohatostí hudební složky.

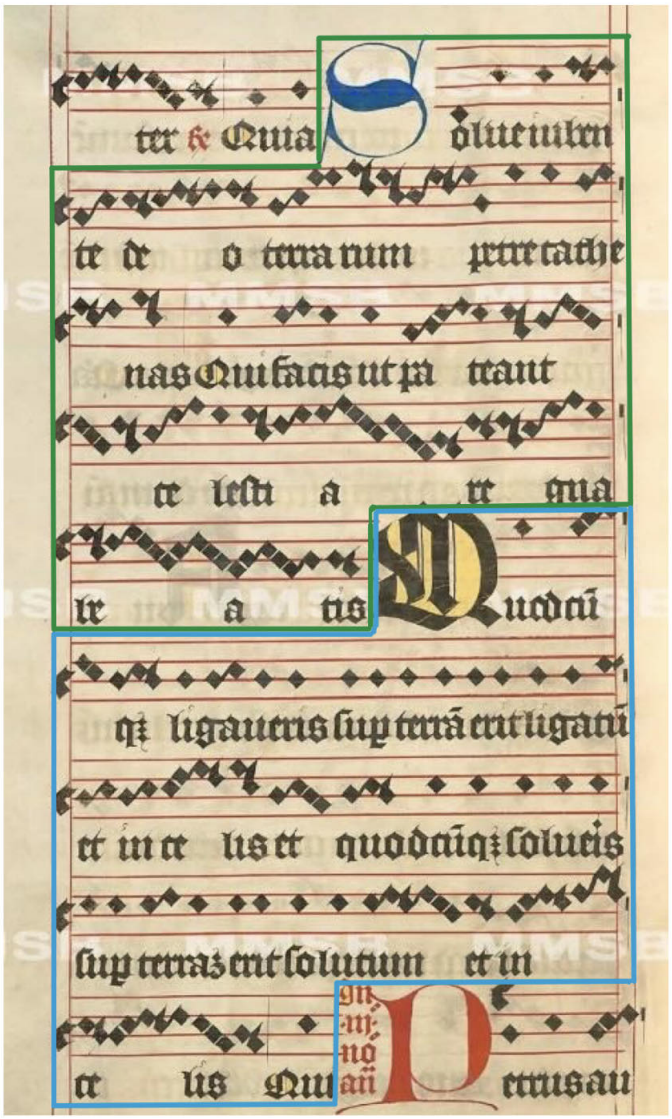
Základním rysem liturgického recitativu je zpěv textu na jednom tónu, přičemž konce vět a jiných úseků jsou zdůrazněny pomocí drobného poklesu či stoupání melodie. Tento postup je tak použitelný pro libovolný text jakéhokoliv rozsahu. Běžně je ho užíváno při provádění modliteb a čtení, tedy u textů, které se téměř každý den roku mění.

Specifický mezistupeň tvoří psalmodie, tedy způsob provádění žalmových textů. Tím, že jsou to texty přednášené se zpěvy volné kompozice v těsném okolí (antifonami či responsorii), došlo k zdokonalení recitačních formulí. Navíc žalmů je omezený počet a používají se všechny v průběhu každého týdne.

Pro volnou kompozici je naopak typické, že každý text má svou, různě složitou, melodii. Jak jsme již zmínili výše, jedná se hlavně o antifony (ty tvoří ohraničení žalmů) a responsoria (ty naopak často ohraničují čtení - lekce), dále pak hymny (zpěv se strofickou strukturou) a invitatoria (zpěvy otevírající noční modlitby oficia). Po raném středověku přibývají ještě sekvence a tropy (žánry volnější a rozmanitější). Mše je rozšířena o žánry odpovídající jejím částem, jež jsou zmíněny v sekci 1.2.1.

[1] Obsáhle o tomto fenoménu hovoří Eliade (2006)

S postupujícími léty a stoletími se zvyšuje složitost repertoáru, píší se pěvecky náročnější kusy většího rozsahu, což vede i k větší rozmanitosti a elaboraci u veršů. Ovšem i tak můžeme v obrázku 1.1 vidět patrný rozdíl mezi hudební složitostí responsoria a verše.



Obrázek 1.1: Volná kompozice a recitativ (Lacoste a kol.)

1.2 Mše a oficium

Liturgický repertoár (tedy všechny části gregoriánského chorálu) můžeme rozdělit na dvě základní části: zpěvy ke mši a zbylé zpěvy, tedy ty používané k tzv. oficiu.

1.2.1 Mše

Mše je pro církevní komunitu vrcholem dne (mše nedělní pak vrcholem týdne), její repertoár je, v rámci jednohlasého chorálu, hudebně nejbohatší. Její zpěvy se dělí

na ordinarium a proprium. Mešní ordinarium zahrnuje neměnné texty, tedy je po každou mši roku stejné. Jedná se o části Kyrie, Gloria, Credo, Sanctus-Benedictus a Agnus Dei. Pěvecky se týká všech přítomných duchovních. Proprium je část, jejíž zpěvy se mění v závislosti na aktuální části liturgického roku, jsou vázané na konkrétní dny či svátky. Patří sem Introit, Graduale, Alleluia, Offertorium a Communio. Nalézáme zde náročnější zpěvy, které jsou prováděny pouze scholou, specializovanou skupinou duchovních, či pověřenou skupinou laiků (jak je tomu obzvlášť dnes).

1.2.2 Officium divinum

Oficium hodinek (lat. *officum divinum*) je mnohem větší a rozmanitější část celého repertoáru řazeného pod gregoriánský chorál. Jedná se o zpěvy určené pro všechny modlitby dne mimo mši, přičemž čtení najdeme pouze v nejrozsáhlejší části zvané Matutinum. I tyto zpěvy mají svůj systém a jednotlivé bloky nastávají v určený čas, proto také název hodinky. V repertoáru oficia nalézáme mnohem větší rozmanitost např. napříč regiony či mezi kláštery a kostely. Právě druhé zmiňované od devátého století přináší dvě formy oficia: monastické, pro mnišské řády, a sekulární, pro zbylé duchovní. Největší rozdíly mezi nimi jsou v počtu antifon a responsorií v Matutinu a v počtu žalmů v nešporách.

Ad vesperas	Nešpory	před soumrakem
Ad completorium	Kompletář	před odpočinkem
Ad matutinas	Matutinum	v noci, před Laudy
Ad laudes	Laudes	před rozřeskem
Ad primam	Prima	za rozbřesku
Missa matutinalis	Ranní mše	(v zimě po Tercii)
Ad terciam	Tercie	v devět
MISSA	MŠE	(v zimě po Sextě)
Ad sextam	Sexta	v poledne
Ad nonam	Nona	ve tři
	O velkých svátcích[a]	
Ad vesperas	Druhé nešpory	
Ad completorium	Kompletář	

Pozn: [a] Ve dnech zvláštního významu se v předvečer následujícího dne místo běžných nešpor zpívají druhé nešpory, které se vztahují k danému svátku.

Tabulka 1.1: Denní cyklus oficia s naznačenou mší (Hiley, 2009)

Pro každý den tak bylo předepsáno mnoho zpěvů. V typických dvaceti čtyřech hodinách zazní minimálně dvacet antifon a šest až osm reponsorií různého rozsahu. Během týdne je také vždy přezpíváno všech 150 zpěvů z knihy žalmů.

Podoba Laud

- pět antifon rámcujících pět žalmů (struktura: antifona - žalm - zopakovaná antifona)
- kapitulum (velice krátké čtení)

- hymnus
- *versiculus* (veršík)
- Zachariášovo kantikum (kantikum - žalmický text vyskytující se v bibli mimo knihu žalmů) a k němu příslušnou antifonu

Nešpory vypadají podobně jako Laudy pouze s použitím kantika zvaného Magnificat. Tzv. malé hodinky (prima, tercie, sexta a nona) mají o něco kratší rozsah. Naopak Matutinum je s dohromady čtrnácti antifonami a dvanácti velkými responsorii (v případě monastického pořádku), která jsou prokládána čteními (lekcemi), ještě rozsáhlejším soustem.

Standardní repertoár týdne je navíc četně narušován všemožnými svátečními událostmi, které přináší jiný, nový repertoár, k nim přímo vázaný. Z toho vychází, že duchovní měl v paměti pro celý rok přes 2000 antifon a 800 responsorií (to bez stovek kusů mešního repertoáru).

1.3 Liturgický rok

Cyklická forma není vlastní jen dnům (pořad oficia a mší je každý den téměř identický) a týdnům (neděle slavnostnějším dnem, odlišena od ostatních), ale také celému roku. Mnoho jeho dní má přiřazený nějaký význam navíc. Díky tomu lze repertoár vnímat po menších úsecích, po jednotlivých svátcích. Lze pak hovořit např. o svatovítských antifonách či o responsoriích k adventním nedělím atp. Každý zpěv má své vymezené místo (či svá místa) a každoročně tak živě připomíná více či méně konkrétní Boží skutek a provází věřícího daným duchovním obdobím.

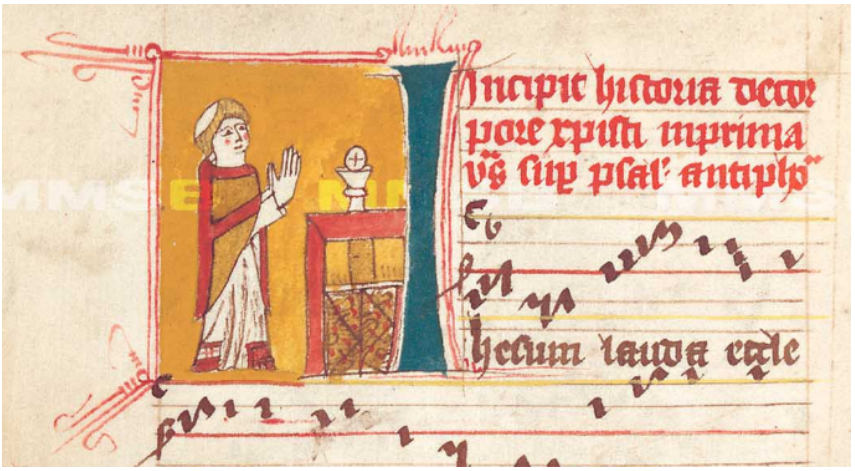
Liturgický rok se sestává ze dvou souběžných cyklů zvaných *Proprium de tempore* a *Proprium Sanctorum*. První zmiňovaný se týká Kristova života a skutků. Je tedy koncentrován kolem Vánoc a Velikonoc. Cykly těchto svátků jsou popsány v tabulkách 1.2 a 1.3. Obě období mezi nimi jsou zvány mezidobím. Obrázek 1.2 ukazuje vyobrazení jednoho z méně známých velikonočních svátků v prameni.

Předpostní doba	tři týdny před postem
Popeleční středa	40 dní před Velikonoční nedělí
Postní neděle (první až pátá)	
Květná neděle	týden před Velikonoční nedělí
Zelený čtvrtek	
Velký pátek	
Bílá sobota	
Velikonoční neděle	
Nanebevstoupení Páně	40 dní po Velikonoční neděli
Letnice (Svatodušní svátky)	50 dní po Velikonoční neděli
Slavnost Nejsvětější Trojice	týden po Letnicích
Slavnost Těla a Krve Páně (Boží tělo)	čtvrtek po Trojici

Tabulka 1.2: Zjednodušený Velikonoční cyklus (Hiley, 2009)

První adventní neděle	neděle nejblíže ke sv. Ondřeji (30. 11.)
Druhá až čtvrtá adventní neděle	
Štědrý den	24. 12.
Boží hod Vánoční	25. 12.
Nový rok (Obřezání Páně)	1. 1.
svátek Tří králů (Zjevení Páně)	6. 1.

Tabulka 1.3: Zjednodušený cyklus kolem Vánoc (Hiley, 2009)

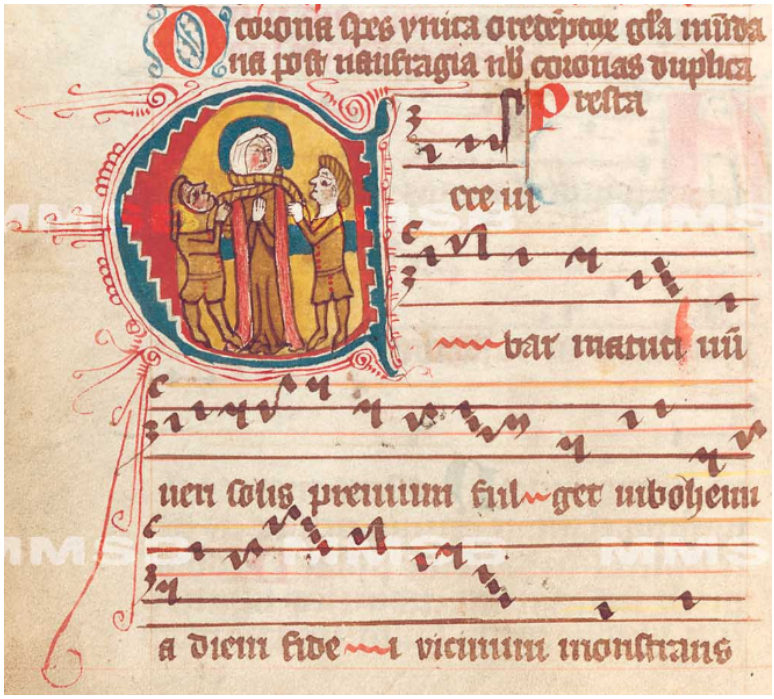


Obrázek 1.2: Ukázka počáteční iniciály k Božímu tělu; Cz-Pu VI G 3a, f. 96v; (Lacoste a kol.)

Druhý cyklus do tohoto pořádku přináší svátky jednotlivých svatých. Jedná se převážně připomínky konkrétních světců (sv. Václava na konci září, sv. Anny v červenci), významných událostí (Nalezení sv. Kříže) či důležitých předmětů (svátek trnové koruny, svátek svatého kopí). Jejich výskyt již není plošný, ale odráží regionální či řádové zvyky. Zřetelně je to vidět např. na českém světci Václavovi. Repertoár k jeho zářijovému svátku nalézáme převážně v českých pramenech, dále pak v několika polských, rakouských a slovenských. Ovšem Francie, Itálie či Španělsko o něm, poměrně pochopitelně, mlčí. Velkou skupinu spadající do části *Proprium Sanctorum* tvoří mariánské svátky, jak naznačuje i tabulka 1.4. Obrázek 1.3 ukazuje možnou iluminaci začátku svátku v prameni.

Uvedení Páně do chámu	*Purificatio Mariae*	2. února (Hromnice)
Zvěstování Panny Marie	*Annunciatio Mariae*	25. března
Jiří	*Georgii*	23. dubna
Jan Křtitel	*Joannis Baptistae*	24. června
Nanebevzetí Panny Marie	*Assumptio Mariae*	15. srpna
Narození Panny Marie	*Nativitas Mariae*	8. září
archanděl Michael	*Michaelis*	29. září
Martin	*Martini*	11. listopadu
Štěpán	*Stephani*	26. prosince
Jan Evangelista	*Joannis Evangelista*	27. prosince

Tabulka 1.4: Přehled významných svátků



Obrázek 1.3: Ukázka počáteční iniciály k svátku sv. Ludmily; Cz-Pu VI G 3a, f. 67v; (Lacoste a kol.)

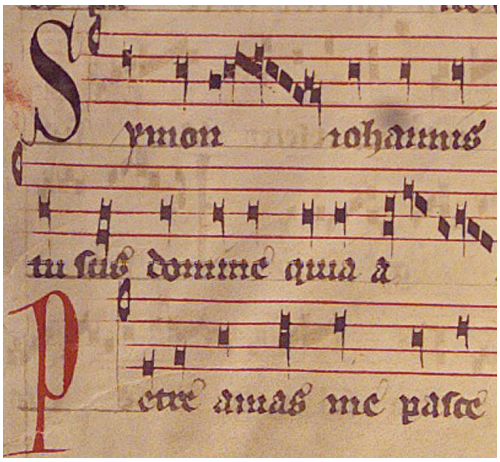
1.4 Prameny

K zápisům zpěvů začalo docházet nejpozději v devátém století, pravděpodobně i dříve. Repertoár nabyl na šíři (nové svátky, nové zpěvy) a také na složitosti (novější kusy často vyžadují větší pěveckou obratnost) a nebylo již možné uchovávat jej pouze jako ústní tradici. Vyskytují se jednak starší nenotované zápisy, které kantorovi připomínají, která antifona s jakým textem má při které příležitosti zaznít, a později i zápisy notované, které osvěžují (a ve fázi rozvinutější notace, tedy od zavedení notové osnovy v 11. století, možná i učí) také odpovídající melodii. Oboje pak pomáhá i snazšímu šíření nového repertoáru. V obrázcích 1.4 a 1.5 můžeme vidět ukázky dvou různých notací.

Prameny v kontextu tradice gregoriánského chorálu jsou liturgické knihy. Ty jsou běžně sestavené ze zpěvů, modliteb a čtení, které jsou organizovány do sekcí odpovídajících částem liturgie. Jednotlivé dny liturgického cyklu mají pro každou část oficia či mše (obzvlášť mešního propria) přiřazeny příslušné texty, které mají ve vhodné formě zaznít.

Jedná se zejména o breviáře a antifonáře (pro oficium) a graduály (pro mši). Později se knihy dělí ještě specifičtěji a narážíme tak na žaltáře (obsahující hlavně žalmové texty a jejich nápěvy), tonáře (pro zápis standardních nápěvů), lekcionáře (na čtení) či sakramentáře (na mešní modlitby). Knihy také můžeme rozdělit podle cyklů liturgického roku: sanctorály (část o světcích a svátcích) a temporály. Repertoár byl, zvláště ve starších pramenech, řazen chronologicky v pořadí liturgického roku, což mimo jiné usnadňovalo orientaci ve velkých svazcích.

Obrázek 1.4: Neumová notace (Lacoste a kol.)

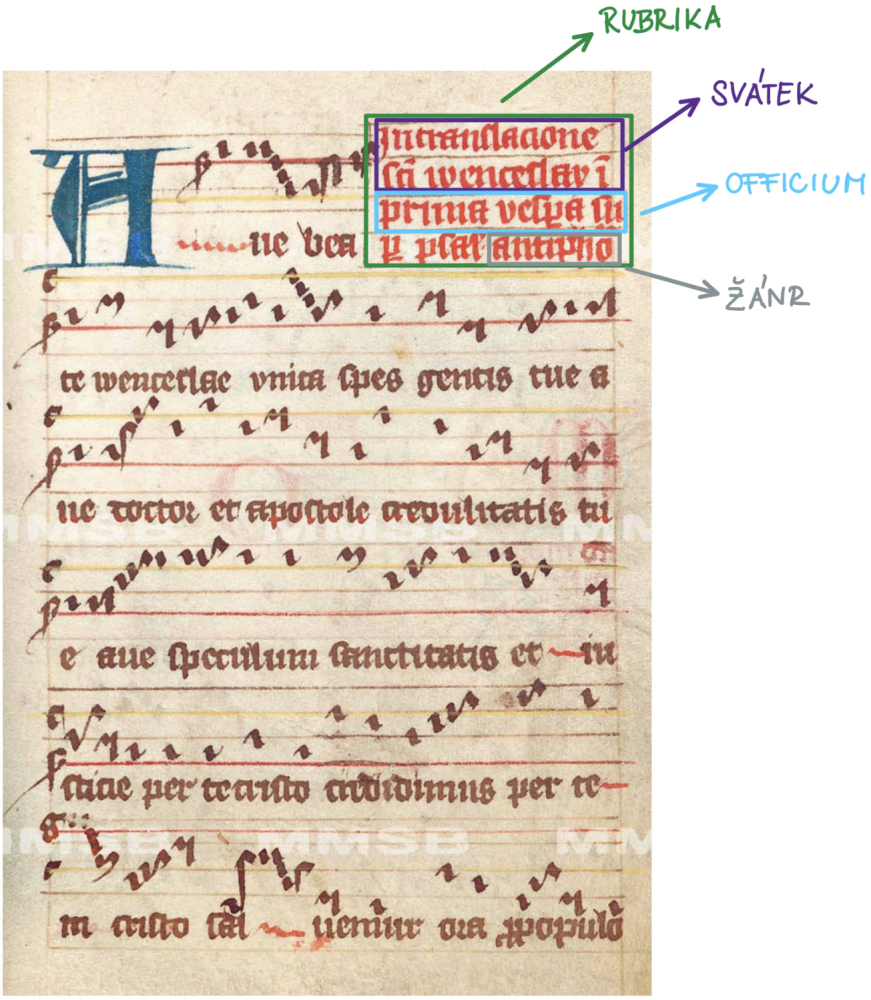


Obrázek 1.5: Kvadratická notace (Lacoste a kol.)

S postupným rozšiřování používaného repertoáru, obzvlášť na základě zavádění nových svátků (jako například svátek Navštívení Panny Marie ve 14. století), dochází k rozšiřování knih o různé dodatky. Zároveň je repertoáru více než knihařská technologie středověku uměla pojmout, proto se setkáváme s liturgickými knihami určenými pro letní nebo naopak zimní část roku. Podobně může být užitečné rozlišovat knihy z klášterů (monastický ritus) a z ostatních církevních jednotek (sekulární ritus).

Na obrázku 1.6 vidíme ukázku z českého pramene z ženského benediktinského kláštera sv. Jiří na Pražském hradě. Jedná se o liturgickou knihu z přelomu 14. a 15. století. Obsahuje směs svátečních oficií a je psaná tzv. rhombickou notací, typickou pro české země. Červené části se říká rubrika a sděluje nám, s jakým kusem repertoáru z pohledu žánru (antifona), oficia (první nešpory) a v tomto případě i svátku (Translace[2] svatého Václava) máme tu čest. Jedná se o základní orientační vodítko v pramenech.

[2]přenesení ostatků, svátek související s kultem relikvií



Obrázek 1.6: Ukázka z pramene Cz-Pu VI G 3a, f. 95r; (Lacoste a kol.)

1.5 Variabilita a přenos

Jakkoliv byly chorální zpěvy považovány za neměnné (potažmo nedotknutelné, jednalo se přeci o kusy svatého slova), nalézají muzikologové napříč Evropou mnohé rozdíly. Ty mohou mít rozmanité příčiny a jedná se tak o téma opakovaně otevírané. Kudy docházelo k šíření? Šel repertoár klášterů jinudy než ten kapitulní či farní? Přenos původních kusů liturgie versus nových svátků pozdního středověku? Jaká byla distribuce svátků v Evropě? Co ovlivňovalo konkrétní světeckou popularitu? Jsou to věci mnišského řádu? Národní? Či úplně lokální? Narážíme při pozorování variability melodií na limitaci lidské paměti? Od každého trochu? Je překvapivé, že máme k dispozici tolik otázek, když mluvíme o liturgii katolické církve, instituci prosazující jednotnost, instituci konzervativní.

I v hudební složce bohoslužeb se vyšlo ze společného základu, ten se ovšem lokálně adaptoval. To se trochu čistí, když Karel I. Veliký po roce 800 n. l. prosadil sjednocení liturgie ve své rozlehlé říši a zároveň když ne o tolik později přichází Benedikt s mnišskou řeholí (to jest návod či soupis pravidel), kterou jsou komunity mnichů nuceny přijmout (alternativou bylo přijmout pravidla života kapituly).

Ovšem jak zmiňuje Hiley (2009), v dalších stoletích mohou kultivační či rozšiřující snahy být odvislé od konkrétní instituce. Například cisterciáci měli od 12. století jasnou představu o tom, jak má liturgie vypadat, a dbali na jednotné provádění napříč svými kláštery. To, že i později zcela jistě docházelo k dobarvování repertoáru, ostatně dosvědčuje také snaha církve v 16. století chorál znovu přísně standardizovat. Přesto, a možná i právě proto, je v těchto směrech dodnes na dochovaném materiálu co zkoumat.

Výše představené by nás mohlo vést k úvaze, že by napříč Evropou mohly existovat skupiny míst, kam např. přišel chorál podobným způsobem v podobné či stejné verzi a kde se tedy používaly převážně shodné zpěvy. Na to lze nahlížet optikou celého repertoáru nebo také pohledem zúženým na jednotlivé svátky a jim příslušející kusy, neboť penzum používaných zpěvů je rozšiřováno postupně, tedy například v různých politických uspořádáních, a také s různou ochotou. Takové skupiny podobných částí repertoáru nazýváme repertoárové tradice.

2. Související práce

Podíváme-li se na výzkum v oblasti přenosu chorálního repertoáru, neomylně musíme narazit na jeden z prvních takových projektů (a zároveň projekt největší a nejproduktivnější) a sice na dlouholetou práci mnichů z francouzského kláštera v Solesmes.[1] Jejich cílem byla pečlivá rekonstrukce a znovuoživení původní podoby mešního i hodinkového repertoáru, včetně podrobné filologické práce. Jde tedy o sledování transmise proti proudu času, zpátky ke kořenům. Výstupem je rozsáhlá edice moderních liturgických knih, jejíž vydávání probíhá od konce 19. století (*Liber Usualis* a později například *Graduale Romanum* či *Antiphonale Monasticum*) do současnosti (revize starších svazků) a která je vnímána závazně.[2]

U této edice ovšem vědecké snahy zkoumat možné tradice a cesty šíření zpěvů zdaleka nekončí a u jejich melodií vlastně teprve začínají. Důležitou součástí všech úvah je debata o ústní tradici a zároveň textových zápisech a jejich možnostech, jak se v hojné míře děje u Treitlera (Treitler, 1981), v jeho shrnující práci, i u Hornby (2004), když rozebírá předchozí dlouholetý trend ve studiu transmise chorálu. Téma šíření ovšem zajímá i De Coula (De Coul, 2021), když zkoumá, jak mnišský řád kartuziánů přijímá nový svátek specificky vzhledem ke své jinak běžné rezervovanosti k novotám. Ottosen (2008), který navazuje na solemské mnichy, přerovnává jejich adventní responsoria a přidává, do velké podrobnosti rozebraná, responsoria k Svátku věrných zesnulých (v českém prostředí také „dušičky"), kde dokonce při výzkumu využívá počítače. Ve všech těchto případech je však studován pouze malý kousek dochovaného liturgického bohatství, ovšem pod pečlivým drobnohledem.

Podhoubí pro digitální gregorianistiku přináší ve velké míře projekt Cantus Database[3] pod vedením Debry Lacoste a Jennifer Bain. Jedná se o zásadní projekt v digitální inventarizaci chorálního repertoáru (část i s melodiemi) a základ dat pro strojový výzkum. To celé ještě rozšiřuje navazující síť databází Cantus Index,[4] která spojuje a sjednocuje rozhraní 19 menším databázím (včetně největší Cantus Database). O přínosu tohoto zdroje dat píše Lacoste (2022).

Existence velkého množství inventarizovaného materiálu bylo využito např. při výzkumu melodické složky repertoáru, o čemž svědčí práce Cornelissena, Zuidemy a Burgoyneho (Cornelissen a kol., 2020a), kteří používají mj. melodie z Cantus Database poskládané do snadno použitelného datasetu CantusCorpus.[5] Kromě knihovny pro Python, pojmenované chant21, která práci s melodiemi v různých formátech usnadňuje, předkládají také dvě případové studie o melodických vzorcích a dále v druhém příspěvku řeší strojové určování melodické modality (viz Cornelissen a kol., 2020b), což je problém pro muzikology dlouhodobě přitažlivý. Na příbuzné téma vznikla také práce Lanze (Lanz, 2023), která s pomocí statistiky řeší vnitřní členění antifon a responosorií a vztah k melodii a modalitě.

Dále došlo k využití sesbírané sady např. v bakalářské práci Kristíny Szabové (Szabová, 2021) a ve v rámci ní vyvinutém nástroji ChantLab,[6] kde jsou

[1] `https://www.solesmes.com/` [cit. 29. 4. 2024]

[2] `https://en.wikipedia.org/wiki/Inter_pastoralis_officii_sollicitudines`?

[3] `https://cantusdatabase.org/` [cit. 10. 4. 2024]

[4] `https://cantusindex.org/` [cit. 10. 4. 2024]

[5] `https://github.com/bacor/cantuscorpus`

[6] `http://chantlab.mua.cas.cz/chants` [cit. 10. 4. 2024]

do výzkumu chorální melodiky zapojeny bioinformatické metody pro *multiple sequence alignment*. Tuto práci rozšiřuje dílo Eiperta s Mossem (Eipert a Moss, 2023b), které přináší knihovnu implementující mj. jednu z představených metod. Jak předvedli Hajič jr. a kol. (2023), podobný postup, v jejich případě zapojení fylogenetických stromů z bioinformatiky, aplikovaný na specifický kus repertoáru, může podat smysluplný výsledek.

Ovšem inventarizovaných dat k pohledu na výskyt repertoáru napříč Evropou je násobně více než dat melodických. Například Cantus Index, jako největší digitální shromaždiště chorálních kusů, k únoru 2024 zpřístupňoval celkem 825 851 záznamů o repertoárových kusech, ovšem pouze 58 357 z těchto záznamů bylo zinventarizováno včetně melodie.[7] K jejich zkoumání existuje nástroj Cantus Analysis Tool[8] operující nad částí sítě databází Cantus Index. Ten vyhodnocuje tradice přístupem vyžadujícím kompletní průnik zpěvů a umožňuje vždy analýzu nanejvýš jednoho svátku. Navíc jeho složka vizualizující tradice neposkytuje uspokojivé výsledky s jasnou interpretací. V minulosti nad Cantus Database ještě běžel nástroj na stavbu dendrogramů hierarchickou shlukovou analýzou a to nad sériemi responsorií, jak o tom píše Lacoste (2012). Nástroj v dnešní době již není dostupný.

Z nečekaně mála počinů na poli přenosu repertoáru v digitálním podání jmenujme Eiperta s Mossem, kteří ve svém příspěvku Communities in Medieval Troper Networks are Shaped by Carolingian Politics (Eipert a Moss, 2023a) na specifickém korpusu zpěvů (tropů) pomocí algoritmu Louvein na detekci komunit ukazují možnou souvislost mezi jejich distribucí po Evropě a třemi částmi Karolínské říše, jak vznikly v 9. století.

Zde ještě stojí za zmínku článek od Le Bomina, Lecointrea a Heyera (Le Bomin a kol., 2016), kteří pracují se sesbíranou hudební tradicí v Gabonu, jejíž podstatná komponenta je také ústní přenos. Řeší téma vertikálního versus horizontálního přenosu hudebního materiálu a zároveň ukazují, že konzistence na velkém prostoru je v takovém případě problematická. Podobným směrem míří i Savage (2019), když upozorňuje na užitečnost etnomuzikologického pohledu na tradici sice geograficky „naši“, ale vzdálenou v čase a socioekonomickém i kulturním prostředí natolik, že ji vlastně lze považovat za cizí.

Zkušenosti s použitím sítí na středověkou literaturu (tedy dochované knihy) předkládá Fernández Riva (2019). Jeho výzkumným cílem je fenomén svazování více různých textů do jedné knihy a sledování přenosu a zároveň celková ukázka práce s analýzou sítí na literárním materiálu, včetně kolekce dat. Podobné téma zpracoval i Vozár (2018), jehož zajímala česká reformační díla, tedy distribuce autorů v nich, a sběr metadat z portálu Manuscriptorium.

Z výše představeného se jeví, že většina dosavadního výzkumu ve výpočetní gregorianistice se týkala melodií, nikoliv repertoáru, přestože katalogových záznamů pro výpočetní výzkum chorálního repertoáru je k dispozici řádově více. Badatelské snahy se zatím omezovaly na konkrétní výseky repertoáru s patřičným detailem či na melodickou složku chorálu. Vnímáme zde tedy prostor k výpočetnímu výzkumu struktury a variability chorálního repertoáru jako celku, tak nějak z ptačího pohledu.

[7] údaj vychází z datové sady, jak je popsána v kapitole 3

[8] `https://cantusindex.org/analyse` [cit. 10. 4. 2024]

3. Datová sada

Vhodná datová sada pro (nejenom výpočetní) výzkum gregoriánského chorálu je už víc jak dvacet let kolektivně shromažďována díky databázi Cantus a síti databází Cantus Index (viz Lacoste (2022), která spojuje 19 projektů (často národních) se stejným rozhraním. Máme tak k dispozici snadný přístup k práci mnoha anotátorů napříč (nejen) Evropou.

Zpěvy jsou pro nás základní jednotkou informace, což dobře lze díky unikátním identifikátorům zvaným Cantus ID (CID), kde každé takové ID zastupuje jeden zpěv, tedy jeden prvek repertoáru (množiny všech používaných zpěvů). Záznamů o jeho výskytu pak může být libovolně mnoho, kromě knihy původu nás o zpěvech zajímá také například v jaké liturgické části oficia byly použity (v prvních nešporách) či tzv. incipt (prvních pár slov textu). Zaznamenaných zpěvů máme ke 400 000. Druhým důležitým prvkem jsou liturgické knihy (prameny), mezi jejichž atributy patří mj. signatura,[1] století vzniku či místo původu. Knih s dostatečným počtem záznamů je v datové sadě 250. Možnost dělit repertoár na menší logické části nám dává vazba jednotlivých zpěvů na svátky (dny roku), při kterých byly zpívány.

Datovou sadu, jejímž základem jsou záznamy o jednotlivých zpěvech v liturgických knihách získané z Cantus Index, laskavě poskytl projekt DACT.[2] Jedná se o verzi dat z databází přístupných skrze Cantus Index z února 2024. Používané CSV soubory jsou k dispozici v githubovém repozitáři této práce[3] či v elektronické příloze ve složce *data*.

3.1 Organizace souborů

Výzkumná část práce (kapitola 4) i implementovaný nástroj (kap. 5) používají jako datové zdroje sedm souborů ve formátu CSV. Dva z nich obsahují informace o zpěvech, jeden o svátcích, a zbylé čtyři se vzájemně doplňují v popisu jednotlivých pramenů.

CSV je jedním z nejčastějších datových formátů používaných pro uložení tabulkových dat. Anglická zkratka *comma-separated values* (přeložitelná jako *čárkami oddělené hodnoty*) odkazuje na standardní způsob oddělování jednotlivých sloupců reprezentované tabulky. Používání jiné interpunkce (např. středníku) je také možné a vyskytuje se obzvlášť v datech, která sama obsahují čárky jako součást ukládané informace. Rozdělováním s pomocí jiného znaku se vyhneme nutnosti používat únikové znaky (anglicky *escape characters*). Řádky odděluje jednoduše odřádkování. Data jsou uložena v prostém textu (anglicky *plaintext*), což je činí snadno zpracovatelnými.

Pro práci s CSV má Python několik knihoven. V této práci používáme pro čtení příslušných souborů modul *pandas*[4], neboť si umí dobře poradit s různými nestandardními situacemi, jakými mohou být chybějící data, odřádkování v rámci

[1] kód specifikující umístění ve sbírce

[2] `https://dact-chant.ca/`

[3] `https://github.com/DvorakovaA/Mapping_the_Repertoire_of_Gregorian_Chant/tree/main/thesis/data`

[4] `https://pandas.pydata.org/`

buňky (tedy jinde než na konci řádku po správném počtu oddělovacích znaků) či podivné znaky aj. K volbě CSV vedla existující zavedená specifikace pro data pocházející z databází Cantus, jak ji představili Cornelissen a kol. (2020a).

3.1.1 Zpěvy

Databáze v Cantus Index jsou organizovány pomocí unikátních identifikátorů Cantus ID. Způsob použití připouští drobné textové varianty v rámci zpěvu jednoho ID. Naopak je striktně odmítavé k užití stejného ID u stejného textu, který se vyskytuje v různých žánrech. (Stejný kus biblického textu může často být použit jako verš k responsoriu v jednom prameni či svátku a jinde jako samostatný *versiculus* či ještě nějak jinak.) Cantus ID naopak není definované s ohledem na svátky či liturgické pozice: antifona pod jedním ID může být použita např. v různé dny Velikonoc či v různých částech oficia.

Hlavní datové soubory pro výzkum i nástroj jsou *all-ci-antiphons.csv* a *all-ci-responsories.csv*. Obsahují informace o záznamech všech antifon, respektive responsorií, které byly v Cantus Index dostupné v únoru 2024. Strukturu souborů, tedy jednotlivé jejich sloupce, popisuje tabulka 3.1, jejíž struktura je přímo odvozená z menší datové sady CantusCorpus (viz Cornelissen a kol. (2020a)). Řada zpěvů může mít některé položky nevyplněné, což ovšem při dodržení potřebného minima informací nebrání použití záznamů.

3.1.2 Prameny

Jednotlivé prameny identifikujeme na základě URL adresy, na které leží jejich podrobnější popisy (viz *drupal_path* v tabulce 3.1). To zároveň jednoduchým způsobem zajišťuje unikátnost takového identifikátoru. Vedle toho je možné, pro snazší orientaci lidí, knihy reprezentovat také pomocí signatur. Přesná podoba souboru je popsána v tabulce 3.2.

Používáme ještě dva soubory rozšiřující informace o původu knihy o důležitou geografickou složku. Základem identifikátorů provenance (místo původu) u pramenů je sada ID z CantusCorpus, ovšem při aktualizaci dat jsme byli nuceni ji rozšířit o nové identifikátory, které na CantusCorpus plynule navazují. Součástí práce bylo obohacení provenancí o geografické údaje, pro které je pravě identifikátor provenance primárním klíčem. Toto najdeme v souboru *geography_data.csv*, jehož struktura je v tabulce 3.3.

Setkali jsme se ovšem s nestandardizovaným užíváním položky provenance napříč jednotlivými databázemi, které jsou v Cantus Index spojené. Jedno místo tak lze najít pod několika různými názvy (anglická vs. národní pojmenování, otázníky v názvech a podobně). Abychom mohli zachovat původní názvy zvolené autory a zároveň se zbytečně nerozcházeli s databázemi, vytvořili jsme soubor *provenance_ids.csv*, který poskytuje převod mezi pojmenováními získanými z webů jednotlivých databází a konkrétní geografickou reprezentací (viz tabulka 3.4).

Následně jsme ze souboru všech pramenů vytvořili soubor *sources-with-provenance-ids-and-two-centuries.csv* obsahující pouze větší prameny (to jest ty mající sto a více záznamů antifon či responsorií) a z původního souboru zachovávající ty sloupce, které jsou následně aplikací či při experimentech používány. Dále s pomocí *provenance_ids.csv* došlo k namapování názvů provenance (sloupec *prove-*

nance z tabulky 3.2) na ID. Zároveň jsme soubor rozšířili o sloupec *num_century*, založený na hodnotách sloupce century, který sice nepřináší žádnou novou informaci, ale usnadňuje výpočetní práci a opět eliminuje nejednotnost v databázích (např. *14th century* ale také *around 1350*). Postup popsaný v tomto odstavci lze najít a replikovat s pomocí skriptu *new_csv.py* v programovacím jazyce Python umístěném v elektronické příloze práce. I zde platí, že ne všechny položky musí nutně být vyplněny.

Položka	**Popis**
id	automaticky generované ID v databázi
corpus_id	okem čitelné ID identifikující zpěv v CantusCorpus
incipit	incipit (prvních několik slov) zpěvu
cantus_id	Cantus ID pro identifikaci v Cantus Index
mode	modus zpěvu
finalis	finála (poslední nota) zpěvu
differentia	melodické zakončení žalmů
siglum	signatura manuskriptu, ze kterého je zpěv
position	liturgická role zpěvu (např. třetí antifona v Laudes)
folio	strana v manuskriptu, kde je zpěv zapsán
sequence	pořadí zpěvu na stránce
marginalia	dodatečné upřesnění umístění zpěvu
cao_coordinates	reference na starší literaturu
feast_id	svátek, při jehož příležitosti se zpěv používal
genre_id	žánr zpěvu (např. responsorium)
office_id	část oficia, ve které se zpěv používá (např. v nešporách)
source_id	id knihy, ve které je zpěv zapsán
melody_id	id melodie, pod kterým je dohledatelná v Cantus Index
drupal_path	URL webové stránky pramene na stránkách domovské databáze
full_text	celé znění textu ve standardizovaném pravopise
full_text_manuscript	celé znění textu, jak je psáno v manuskriptu
volpiano	přepis melodie ve formátu *volpiano*
notes	poznámky
dataset_name	název datového zdroje, do kterého zpěv patří
dataset_idx	index datového zdroje, do kterého zpěv patří
image_link	URL, pod kterým lze najít snímek manuskriptu

Tabulka 3.1: Přehled datových položek hlavních souborů; povinné položky jsou zapsány tučně

Položka	Popis
id	automaticky generované ID v databázi
title	název pramene
siglum	signatura pramene
desctiption	rozšiřující popis pramene
rism	signatura pramene ve standardním muzikologickém formátu[5]
date	textová specifikace období vzniku
century	století původu
provenance	místo původu
provenance_detail	upřesňující poznámka k místu původu
segment	indikátor zdrojové databáze *(nepoužito)*
summary	detailnější shrnutí informací o prameni
indexing_notes	podrobnější poznámky k indexaci
liturgical_occasions	liturgické příležitosti, k jakým je v prameni repertoár
indexing_date	datum zpracování
drupal_path	URL webové stránky pramene na stránkách domovské databáze
cursus	informace, je-li pramen monastický či sekulární
image_link	URL, pod kterým lze najít obrazovou podobu pramene *(nepoužito)*

Tabulka 3.2: Přehled datových položek v souboru *sources-of-all-ci-antiphons_OPTIONAL-CENTURY.csv*; povinné položky jsou zapsány tučně

Položka	Popis
provenance_id	geografické ID místa (např. provenance_123)
provenance	název místa pro orientaci v souboru
latitude	zeměpisná šířka v decimálním formátu
longitude	zeměpisná délka v decimálním formátu

Tabulka 3.3: Přehled datových položek v souboru *geography_data.csv*

Položka	Popis
id	automaticky generované ID v databázi
provenance	název místa, jak je v souboru pramenů
provenance_id	geografické ID místa

Tabulka 3.4: Přehled datových položek v souboru *provenance_ids.csv*

[5] `https://rism.info/`

Položka	Popis
id	automaticky generované ID v databázi
drupal_path	URL webové stránky pramene na stránkách domovské databáze
title	název pramene
provenance	místo původu pramene
siglum	signatura pramene
num_century	století vzniku pramene číselně
century	století vzniku pramene slovně
cursus	informace, je-li pramen monastický či sekulární
provenance_id	geografické ID místa

Tabulka 3.5: Přehled datových položek v souboru *sources-with-provenance-ids-and-two-centuries.csv*

3.1.3 Svátky

Jednotlivým svátkům propůjčujeme identifikátory z datasetu CantusCorpus (položka *feast_id* v přehledu 3.1) a zároveň z něj využíváme soubor *feast.csv* (struktura viz tabulka 3.6) k namapování těchto ID na standardní pojmenování svátků zavedené v Cantus Index.

Položka	Popis
id	unikátní id ve formátu *feast_1234*
name	název svátku
description	popis svátku
date	datum slavení
month	měsíc slavení (číselně)
day	den v měsíci, kdy je slaven (číselně)
feast code	identifikátor z Cantus Index
notes	poznámky

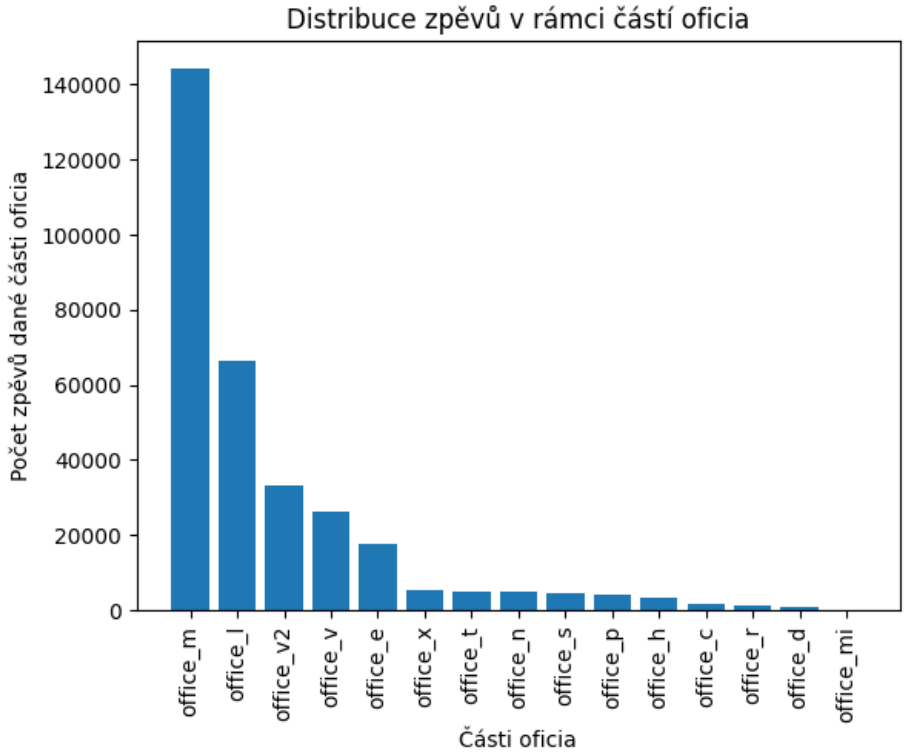
Tabulka 3.6: Přehled datových položek v souboru *feast.csv*

3.2 Přehledové informace o datasetu

Ze všech získaných dat získaných z Cantus Index používá náš výzkum i aplikace toliko jejich podmnožiny. Předně pracujeme pouze se zpěvy z oficia, a to s antifonami a responsorii (ostatní žánry nepoužíváme, neboť jsou obecně mnohem méně pevnými či reprezentativními vzorky). Uvažujeme jenom prameny, a tedy jen zpěvy z nich, které mají sto a více záznamů o antifonách či responsoriích, což je snaha eliminovat fragmenty a vysoce nekompletní záznamy. Zároveň pro vyvíjený nástroj nebereme v potaz záznamy svátků, pro které celkově nemáme, po odstranění nedostatečně velkých pramenů, alespoň pět zpěvů.

Základní údaje o datové sadě

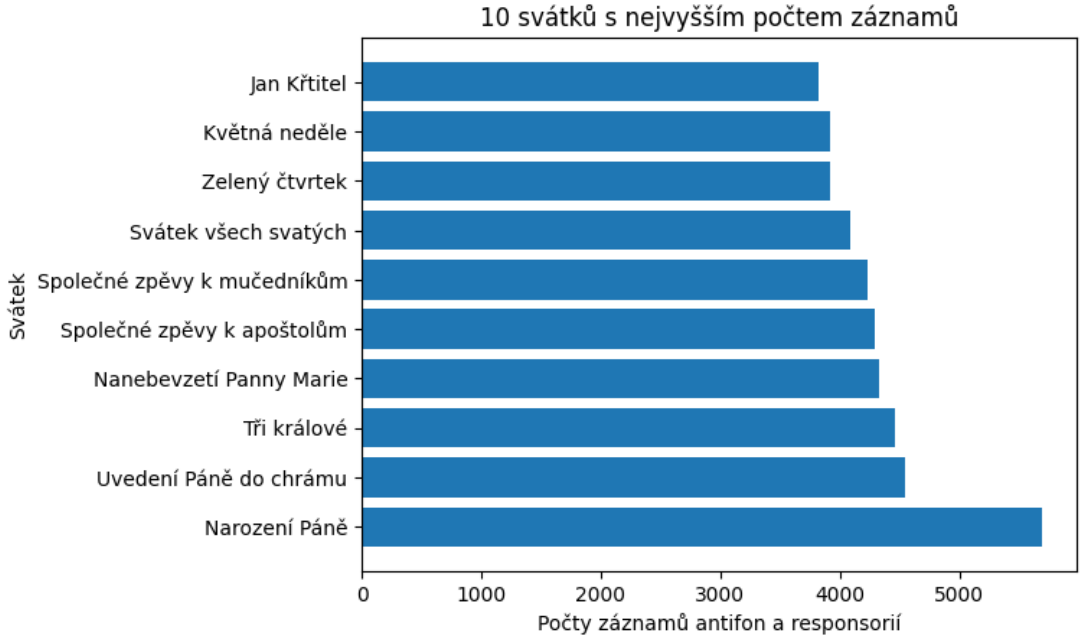
- Záznamů o antifonách celkem: 265 621
- Záznamů o responsoriích celkem: 124 195
- Záznamů o pramenech celkem: 816
- Počet větších pramenů: 250
- Z toho s informací o geografii provenance: 233
- Počet známých geografických lokací: 122
- Záznamů o antifonách velkých pramenů: 242 086
- Záznamů o responsoriích velkých pramenů: 120 546
- Počet Cantus ID ve velkých pramenech: 17 599
- Počet svátků, pro které máme celkově alespoň 5 zpěvů: 1 029
- Záznamů o antifonách velkých pramenů a svátků: 234 355
- Záznamů o responsoriích velkých pramenů a svátků: 116 252
- Počet Cantus ID ve velkých pramenech a svátcích: 17 144



Obrázek 3.1: Distribuce zpěvů v rámci částí oficia

Graf 3.1 (význam id pro části ofica lze najít v příloze A.1) nepřináší žádná velká překvapení. Matutinum obsadilo první příčku právoplatně, jedná se o jednoznačně nejrozsáhlejší hodinku z ofica (až devět či dvanáct responsorií a až devět či třináct antifon). Malé hodinky (to jest prima, tercie, sexta a nona) mají, jak už název napovídá, obecně malý rozsah (jedna antifona a nanejvýš jedno responsorium) a zároveň jsou často těmi částmi, pro které se speciální oficium nepřiřazuje (a v daný moment se použije repertoár z běžného, nesvátečního, dne). Kompletář je téměř nevariabilní částí dne a proto také v knihách minimálně uváděnou.

Laudy (ranní oficium) a nešpory (večerní oficium) bychom, vzhledem k jejich velice podobné struktuře (pět či šest antifon), možná očekávali početně blíže k sobě, než ukazuje graf 3.1. Toto rozložení může být podpořeno dvojicí faktorů. Laudy jsou obecně více propriální, to jest jsou často tou částí oficia, pro kterou je připraven speciální repertoár vztahující se k danému dni (svátku). Naopak do nešpor se *proprium* propisuje až v pozdějším středověku, mnohem častěji se tedy berou zpěvy k nim ze standardního žaltáře a nejsou v zápisech o svátečních dnech znovu. Možnou druhou příčinou by mohl být fakt, že pokud se v daném dni používal stejná množina zpěvů pro první a druhé nešpory (což nebylo tak neobvyklé, někdy se zpěvy sdílely i s laudami), pak se napsala pouze rubrika,[6] že se tak má stát. A při inventarizaci nedochází k opisům záznamů. Převaha druhých nešpor nad prvními je pak odůvodnitelná příčinou jejich výskytu. Jedná se o hodinku, která se objevuje při větších svátcích, což je ovšem přesně to, co zaznamenáváme do liturgických knih – právě ty zpěvy, které nepotřebujeme denně a je vhodné je mít zaznamenané k osvěžení.

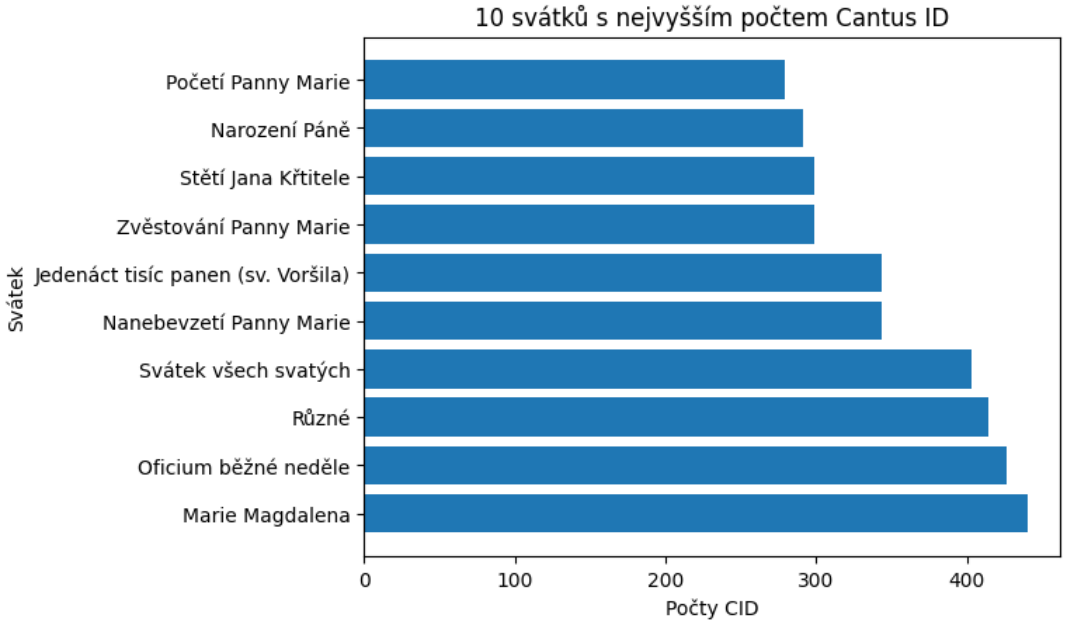


Obrázek 3.2: Svátky s největším počtem záznamů antifon a responosorií

Grafu 3.2, který odhaluje svátky, k nimž máme nejvíce záznamů o používaných antifonách a responsoriích, vévodí Narození Páně, vrchol celosvětově asi nejvýraznějšího křesťanského svátku – Vánoc. V ne zcela těsném závěsu je tříkrá-

[6] poznámka usnadňující orientaci v knize, ukázka viz obrázek1.6

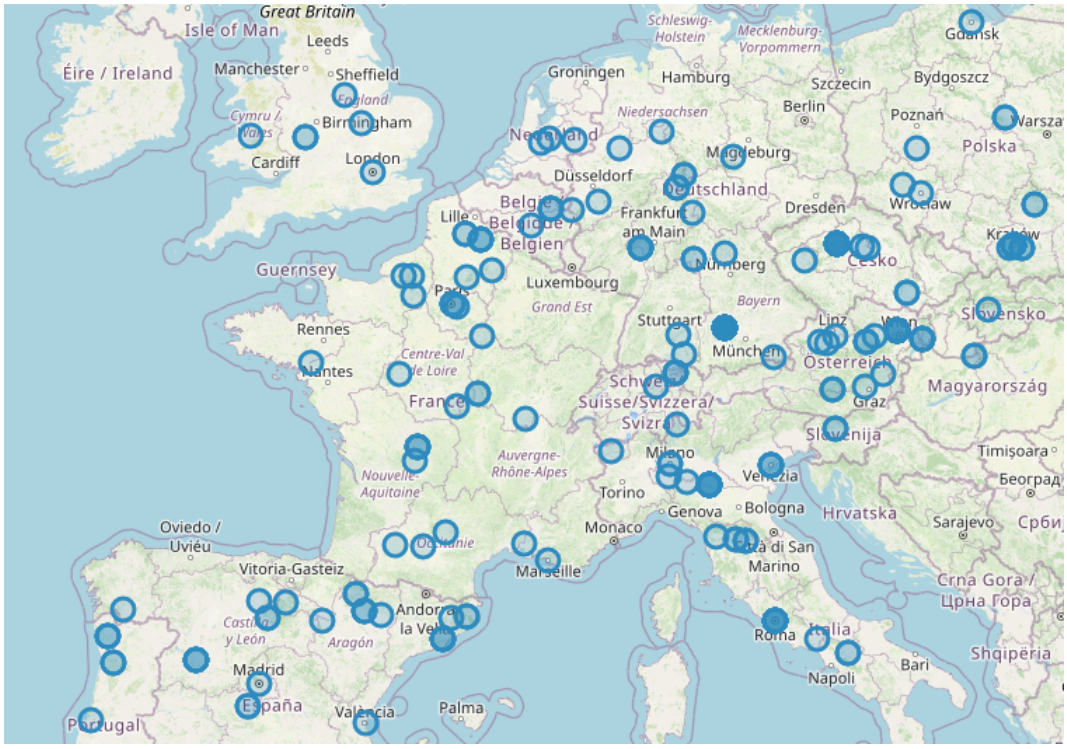
lový dozvuk vánočních svátků a také dva mariánské svátky.[7] Dále tzv. *Commune*, v grafech pojmenované jako společné zpěvy, odráží realitu liturgických knih, kdy v případě např. svátku světce mučedníka z nějakého důvodu ne dostatečně významného, aby byl k dispozici jeho vlastní repertoár, je možné ho oslavit obecnějšími zpěvy týkajícími se právě třeba mučednické smrti. Velikonoce, tedy nejvýznamnější svátek roku, jsou v „první desítce“ zastoupeny Zeleným čtvrtkem a Květnou nedělí, která velikonoční týden otevírá.



Obrázek 3.3: Svátky s největším počtem různých Cantus ID k nim přiřazených

Podobně i z přehledu 3.3, který odráží rozmanitost jednotlivých svátečních částí repertoáru, můžeme, vzhledem ke třem zástupcům, vyčíst popularitu a rozšířenost mariánských svátků. Překvapující naopak může být připomínka popravy 11 000 panen v německém Kolíně, která se svátkem sv. Voršily (či Uršuly), jako nejvýznamnější z nich, trochu splývá, přestože má v rámci Cantus Index svou položku. Obecně by se dalo usuzovat, že svátky z temporálu, tedy ty zaměřené na život a úmrtí Ježíše Krista, nemají takovou variabilitu, tedy tolik různých kusů, jako ty sanktorálové, příslušné světcům a událostem jejich životů. Zástupce *de tempore* je mezi deseti „nejkošatějšími“ položkami pouze jeden.

[7] Uvedení Páně do chrámu je svátek *Purificatio Mariae* – Očištění Panny Marie



Obrázek 3.4: Mapa dostupných geograficky podložených větších pramenů

3.3 Limitace použité datové sady

Při interpretaci výsledků na těchto datech (i obecně na datech o gregoriánském chorálu) je potřeba dát si pozor na nejméně tři věci. První dvě z nich naráží na množství dat, poslední na velikost projektu.

Inventarizace starých rukopisů, byť jich je v dnešní době množství digitalizováno[8], stojí, z vlastní zkušenosti[9], netriviální množství času, zápalu a energie poučených pracovníků, čehož se ne nutně dostává všude, kde se nějaké prameny dochovaly. A i tam, kde se na inventarizaci pracuje, nemusí být vše dostupné již hotové (např. u nás v Praze). To vede k nerovnoměrné distribuci informací v datech, což může výsledky zkreslovat. Zároveň prameny nemají stejnou (uniformní) šanci se do databází dostat, neboť jejich zpracování je často vázáno na různé projekty a rozložení zpracovaného nelze považovat za náhodné. Toto lze pozorovat i v mapě 3.4.

Druhý kámen je tak trochu zakopán v historii. Nejde totiž jenom o to, které prameny se nám již dostaly do databáze, ale také o to, které se vůbec dostaly do současnosti. Aktuální odhady na množství dochovaných písemných pramenů ze středověku se pomocí metod používaných v ekologii uvádí jako 7 až 10 % (viz Kestemont a kol. (2022)). Mnohé nástrahy čekaly katolické knihy nejenom v českých zemích, ale i jinde v Evropě. Také přístup ke svazkům, od jejichž užívání se třeba již upustilo, mohl být různý, od archivačního až po, řekněme, recyklační. A to je mezera, kterou nám nepomůže zalepit žádná mravenčí práce.

Poslední problém už byl drobně nastíněn v sekci 3.1.2. Jelikož Cantus Index sdružuje výsledky řady různých dílčích katalogizačních projektů za několik de-

[8] např. díky `https://www.manuscriptorium.com/`

[9] indexování pod vedením prof. Ebena pro databázi Fontes Cantus Bohemiae

setiletí existence, narážíme na některé nejednotnosti a nedokonalosti. Například *cursus* je údaj, který poměrně často chybí (z 916 pramenů je uveden jen u 242). Je ovšem poměrně pravděpodobné, že ne vždy je to proto, že by byl skutečně neznámý, byť u fragmentů to tak snadno být může, ale je spíš prostě jenom nevyplněný (z textového popisu pramene lze původ vyčíst). Podobně při snaze rekonstruovat přesné liturgické pozice, např. při četbě informace „jedná se o třetí antifonu v druhém Nocturnu Matutina“, jsme narazili na tolik variant zápisu dané informace, že bylo nemožné pozice spolehlivě automatizovaně určit. V popisu grafu 3.1 pak nastiňujeme ještě problém s případnou snahou o rekonstrukci celého dne na základě záznamů v databázích.

To všechno nás vyzývá k opatrnosti při interpretaci výzkumných výsledků. V rámci uvažování nad výběrem dat k výzkumným výpočtům je potřeba přemýšlet nejenom v kontextu toho, co je dostupné v databázích, ale právě i v kontextu toho, co v nich, z různých důvodů, může chybět.

4. Výpočetní hledání tradic

Jedním z polí vědeckého bádání muzikologů v oblasti chorálu a hudby obecně je koncepce tradic. Existují v rámci pramenů repertoárové skupiny, které by vykazovaly nějakou podobnost? Je v šíření zpěvů a oficií napříč západním křesťanským světem patrný nějaký směr či regionální systém? Jedná se snad o věc chronologického vývoje? Jak silnou roli hrají řády, na jejichž možný vliv naráží De Coul (2021), či politická historie, jak naznačují Eipert a Moss (2023a)?

Pro konkrétní části materiálu lze narazit na rozbory repertoáru. Uveďme např. článek Steinerové zaměřující se na lokální a regionální tradice antifon k invitatoriu (Steiner, 1985), práci na oficiu ke Svátku věrných zesnulých od Ottosena (Ottosen, 2008) či samotný *Corpus Antiphonalium Officii*, kde René-Jean Hesbert porovnává obsah dvanácti středověkých antifonářů s cílem s jejich pomocí najít archetyp oficia (Hesbert, 1963–79). Jeho dílo ovšem může sloužit také jako doklad o nejednotnosti a tradicích, jak o tom píše Steinerová (Steiner, 1985). Hiley (1993) zmiňuje vedle Hesberta i další výzkumy zaměřené na porovnávání určitého kousku repertoáru mezi prameny (např. Le Roux (1961), které prokazují schopnost izolovat rodiny příbuzných pramenů např. podle diecéze, klášterních řádů atp. Cílem následující kapitoly je popsat snahy o objevování tradic ve větším měřítku, s použitím velkých kusů repertoáru a s pomocí výpočetních technologií.

4.1 Metody

Domníváme se, že na náš muzikologicky definovaný problém tradic lze očima informatika nahlížet jako na shlukovou analýzu (anglicky *clustering*) či přesněji možná spíš úlohu detekce komunit v síti. Repertoárová síť, jak o ní píše Roy (2022), je v našem případě tvořena jednotlivými inventarizovanými prameny (liturgickými knihami), jejich propojení je pak odvozeno z jejich obsahu – u nás konkrétně ze zapsaných antifon a responsorií, kteréžto lze snadno jednoznačně reprezentovat pomocí jejich Cantus ID. Těmto spojením přidáváme váhu (udáváme vzdálenost mezi nimi), neboť nám nejde o binární vztah (zda sdílí nějaké kusy či nikoliv), ale zajímá nás i míra případného překryvu. To, že považujeme výpočetní metody za aplikovatelné, je dáno tím, že máme k dispozici nezanedbatelné množství dat (viz kapitola 3.2).

4.1.1 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise), viz Ester a kol. (1996), je jednou z metod shlukové analýzy. Funguje na principu hledání shlukových jader s vysokou hustotou, kolem kterých pak staví příslušné komunity (shluky). Jeho výhodou pro náš problém je, že nepotřebuje na vstupu znát počet jader.

Použili jsme implementaci z knihovny *scikit-learn*[1] (viz Pedregosa a kol. (2011)). Ta vychází z algoritmu, jak ho popisuje Ester a kol. (1996). Použitá verze nám umožňuje data předat ve formě matice vzdáleností a pracovat s vyžadovanými

[1] `https://scikit-learn.org/stable/index.html`

dvěma parametry – ε-sousedstvím (`eps`), což je maximální vzdálenost mezi vzorky, ve které jsou ještě považovány za sousedy, a s minimálním počtem sousedů potřebných k tomu, aby prvek mohl být jádrem (`min_samples`).

Pro výpočet vzdálenosti mezi množinami zpěvů z jednotlivých pramenů používáme Jaccardovu vzdálenost (poměr velikosti průniku a sjednocení) a dále také vzdálenost založenou na modelování témat, jejíž výpočet je popsán v sekci 4.1.3. V obou případech musíme ještě získanou hodnotu odečíst od jedničky, aby se jednalo o vyjádření vzdálenosti (úplně stejné množiny mají Jaccardovu vzdálenost 1, ovšem v případě vyjádření vzdálenosti chceme říci, že jsou stejné, tedy od sebe ve vzdálenosti 0).

4.1.2 Louvein algoritmus

Louvein algoritmus[2] na detekci komunit je postupem používaným v *Digital humanities*, jak ukazují například Eipert a Moss (2023a) či zapojení do analýzy sociálních sítí (anglicky *Social network analysis*) u Knyazevy (Knyazeva, 2021). Jedná se o heuristický způsob získání komunit ze síťové struktury. Naše síť má na místě vrcholů jednotlivé prameny. Váhy hranám jsou přidělené na základě zpěvů, které se v knihách vyskytují. Využíváme na to poměr mezi průnikem a sjednocením (Jaccardova vzdálenost), který je symetrický, a jedná se tedy o neorientovanou síť.

V této práci využíváme implementaci z knihovny *networkx*[3] (viz Hagberg a kol. (2008)), která je založená na optimalizaci tzv. modularity, jak je popsána u Blondela (Blondel a kol., 2008). Jedná se o skóre, jehož změny v závislosti na změnách v síti pozorujeme a které chceme maximalizovat.

Zisk modularity při přičlenění vrcholu i do komunity C, lze vyjádřit jako

$$\Delta Q = \frac{k_{i,in}}{2m} - \gamma \frac{\sum_{tot} k_i}{2m^2},$$

kde m je počet vrcholů grafu, $k_{i,in}$ je součet vah hran z i do vrcholů v komunitě C, k_i je součet vah hran z i, $\sum_{tot}$ je součet vah hran, které jsou incidentní vrcholům z C, a γ je parametr rozlišení (rezoluce, `resolution`). Je-li menší než 1, algoritmus preferuje větší komunity, je-li větší, pak naopak.

Algoritmus pracuje ve dvou krocích, které se opakují. V prvním z nich je každý z vrcholů považován za vlastní komunitu. Komunity jsou následně procházeny a pro každou z nich je vypočítána změna modularity, která by nastala při jejich přiřazení ke každé z dalších komunit. Následně dojde k přiřazení zkoumané komunity tam, kde bude celkový přínos největší. Je-li vůči všem zbývajícím zisk modularity záporný, zůstává dál vlastní komunitou. Toto probíhá, dokud existuje změna vedoucí ke zlepšení modularity.

Ve druhém kroku pak dochází k postavení nové sítě. Její vrcholy jsou komunity nalezené v předchozí části a je sestavena tím způsobem, že váhy hran mezi starými vrcholy v každých dvou komunitách jsou sečteny a považovány za váhu nové hrany mezi novými vrcholy vytvořenými z daných komunit. Načež se vracíme ke kroku jedna.

[2] Český překlad by byl pravděpodobně Lovaňský algoritmus, avšak nepůsobí zavedeně.

[3] `https://networkx.org/documentation/stable/index.html`

Algoritmus zastaví při skončení změn v modularitě, při poklesu změn pod zvolený limit či při dosažení nastaveného maximálního počtu kroků. Opakovaný běh algoritmu může vracet různé výsledky, neboť záleží na pořadí, v jakém jsou vrcholy v kroku jedna uvažovány, a toto pořadí je náhodné.

4.1.3 Modelování témat

Jinou formalizaci problému tradic v chorálním repertoáru nabízí počítačová lingvistika. Jednou z jejích úloh je modelování témat (anglicky *topic modeling*), tedy hledání abstraktních témat nad kolekcí dokumentů. Výsledkem je pravděpodobnostní distribuce jednotlivých témat pro každý dokument kolekce (z *document-topic* matice) a také extrakce klíčových slov pro každou skupinu (z *topic-term* matice).

Tento postup jsme aplikovali na náš problém a datovou sadu. Dokumenty jsou představovány liturgickými knihami (prameny). K jednotlivým Cantus ID zpěvů v dané knize zaznamenaným přistupujeme jako k výskytu slova v příslušném dokumentu. Po vyhodnocení lze ze získané pravděpodobnostní distribuce vzít dominantní tematickou skupinu pro každý dokument a na základě ní seskupit dokumenty do skupin (komunit). Zároveň je to ale metoda umožňující, „měkčí“ přístup, který by mohl být adekvátnějším vyjádřením složité struktury repertoáru.

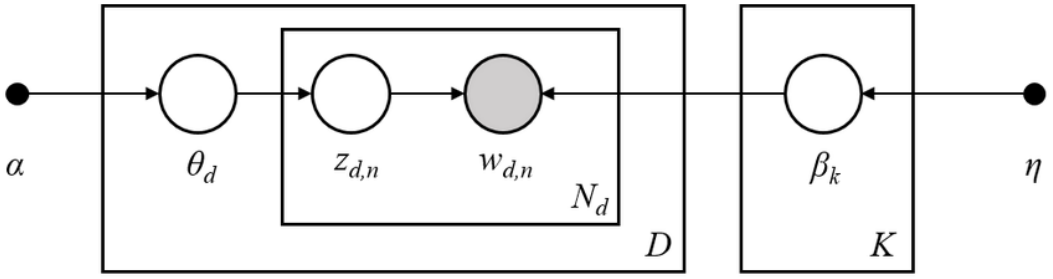
Nastíněná aplikace není po formální stránce problematická, avšak standardně je v počítačové lingvistice pro modelování témat pomocí modelu LDA, jak je popsán vzápětí, užíváno mnohem větších kolekcí. V našem případě mluvíme o 250 dokumentech (pramenech) a po odebrání velice četných a zároveň i příliš specifických zpěvů zůstaneme zhruba s pěti až deseti tisíci položkami ve slovníku (celkem je zpěvů přes sedmnáct tisíc), zatímco lingvisté mívají spíše tisíce dokumentů a desetitisíce slov.

LDA

Pro modelování témat jsme použili implementaci pravděpodobnostního modelu zvaného *Latent Dirichlet Allocation* (LDA) konkrétně z knihovny *scikit-learn*[4] (viz Pedregosa a kol. (2011)). Ta vychází z práce, kterou představili Blei a kol. (2003) a Hoffman a kol. (2010). Na rozdíl od obou zbylých použitých, výše zmiňovaných, algoritmů, zde je třeba předem říci, do kolika témat chceme kolekci dokumentů namodelovat (tedy kolik shluků či komunit chceme vrátit).

Klíčovou vlastností LDA je používání Dirichletovského apriorního rozdělení pro distribuci témat v dokumentu, a to s parametrem $\alpha < 1$ (typicky $\alpha = 1/K$). To vytváří tlak na to, aby každý dokument patřil do co nejméně témat – v našem případě tedy preferujeme, aby pramen patřil co nejvíce do pouze jedné hypotetické tradice. Ponecháváme však prostor pro nejistotu, pramen může patřit do jisté míry i do některých dalších tradic.

[4] `https://scikit-learn.org/stable/index.html`



Obrázek 4.1: Schéma LDA modelu (Pedregosa a kol. (2011) podle Blei a kol. (2003))

Schéma 4.1 zobrazuje generativní model se třemi úrovněmi, kde jednotlivé rámečky reprezentují opakovaný výběr vzorku, a kde:

- α je pravděpodobnost že téma patří dokumentu (doc_topic_prior),
- D je počet dokumentů v kolekci,
- θ_d je kategorická distribuce témat v dokumentu,
- $z_{d,n}$ je n-tá lokální skrytá proměnná d-tého dokumentu vyjadřující distribuci témat dokumentu pro dané slovo
- $w_{d,n}$ je n-té slovo d-tého dokumentu kolekce (pozorovaná porměnná),
- N_d je počet slov v dokumentu d,
- K je počet témat,
- β_k je k-tá globální skrytá proměnná vyjadřující pravděpodobnost příslušnosti slova k tématu k
- η je pravděpodobnost, že slovo patří tématu (topic_word_prior)
- ζ je poměr počtu výskytů slova v tématu vůči celkovému počtu jeho výskytů.

Parametr ζ tedy udává, jaká jsou nejdůležitější slova daného tématu, a θ nám říká, jaká jsou nejdůležitější témata dokumentu.

Při modelování korpusu dochází k následujícímu generativnímu procesu:

1. Pro každé z K témat:
   (a) Výběr $\beta_k \sim Dir(\eta)$
2. Pro každý dokument d z kolekce:
   (a) Výběr $\theta_d \sim Dir(\alpha)$
3. Pro každé slovo i dokumentu d:
   (a) Výběr tématu $z_{di} \sim Multinomial(\theta_d)$
   (b) Výběr pozorovaného slova $w_{ij} \sim Multinomial(\beta_{z_{di}})$

V rámci modelování textových dat je ještě třeba zmínit pojmy *document frequency* a *term frequency*. Jedná se číselné hodnoty charakterizující četnost výskytu slov. *Term frequency* je vztažená vždy k dokumentu a vyjadřuje, kolikrát se slovo (termín) v daném textu vyskytuje, *document frequency* je vztažena k celé kolekci a pro dané slovo udává počet dokumentů, ve kterých je obsaženo.

Model LDA na vstupu předpokládá *document-word matrix*, kde hodnota buňky na i-tém řádku a v j-tém sloupci vyjadřuje, kolikrát se slovo (v našem případě chorální zpěv identifikovaný svým Cantus ID) vyskytlo v daném dokumentu (v našem případě liturgické knize). Na získání těchto matic jsme použili model *CountVectorizer* z knihovny *scikit-learn*, který výpočet provede za nás. Výhodou je také možnost jeho parametrizace ve smyslu míry zachování slov kolekce, tedy možnost práce s příliš četnými či naopak příliš neobvyklými slovy (zpěvy).

LDA pro výpočet vzdálenosti

Získání pravděpodobností distribuce komunit pro jednotlivé prameny z LDA modelu může mít ještě jiné využití, než přímý zisk komunit skrze hledání maximální pravděpodobnosti. Jedná se zároveň o metodu k redukci dimenzionality, namísto množin zpěvů můžeme prameny reprezentovat pomocí pravděpodobnostních distribucí přes daný (nastavitelný) počet témat. Taková vyjádření pak porovnáme pomocí Jensen-Shannonovy vzdálenosti. Tu implementuje knihovna *scipy*.[5] Pro dva vektory pravděpodobností p a q je definovaná jako

$$\sqrt{\frac{D_{KL}(p||m) + D_{KL}(q||m)}{2}},$$

kde D_{KL} je Kullback-Leiblerova divergence pro dva vektory pravděpodobností p, q a $m = (p + q)/2$.

Jedná se o symetrickou vzdálenost, která je navíc omezená na intervalu 0 až 1, což nám umožňuje ji použít jako váhu pro Louvein algoritmus i jako vzdálenost pro DBSCAN.

4.2 Metriky pro evaluaci

Evaluace je problematickou částí práce, neboť na rozdíl od řešení úloh se známými správnými odpověďmi, kde lze jednoduše vyhodnocovat schopnost zvolených metod dobře odhadovat skutečné výsledky, pro toto výzkumné téma odpovědi neexistují.

4.2.1 Evaluace stabilitou

Metody však můžeme ohodnotit nepřímo na základě stability nalezeného řešení. Stabilita pro nás měří míru citlivosti metody na daných datech, tedy odolnost vůči náhodnému šumu, jakým může být pouhá permutace dat. Tento přístup pozorování jednotnosti odpovědí algoritmu lze použít, neboť všechny tři použité metody vykazují nedeterministické chování. Má tedy smysl porovnávat výsledné rozložení komunit s jinými běhy stejného algoritmu při stejném nastavení jeho

[5] `https://docs.scipy.org/doc/scipy/index.html`

parametrů. Jestliže metoda výrazně mění své odpovědi při opakovaném spuštění na stejné datové sadě, těžko lze takové přiřazení chorálních pramenů do tradic považovat za věrohodné či směrodatné. Tento postup nám nedává možnost rozlišit mezi problémem nevhodnosti metody a případnou neexistencí předpokládaných tradic. Lze ho ale použít pro porovnávání metod.

To činíme vždy po dvojicích, a to pomocí tří měřítek.

- Skóre vzájemné informace:

$$MI(V,U) = \sum_{i=1}^{|U|}\sum_{i=1}^{|V|} \frac{|U_i \cap V_j|}{N} \log \frac{N|U_i \cap V_j|}{|U_i||V_j|},$$

  kde N je počet pramenů, $|U_i|$ je počet prvků ve shluku U_i z verze shluků U a stejně tak pro V.

- Jaccard index:

$$JI(U,V) = \frac{N_{11}}{N_{11} + N_{01} + N_{10}},$$

- Rand index:

$$RI(U,V) = \frac{N_{11} + N_{00}}{\binom{N}{2}},$$

  kde N je počet pramenů, N_{11} je počet dvojic pramenů v U i V zařazených do stejného shluku, N_{00} je počet dvojic pramenů zařazených v U i V do různých shluků a N_{10} a N_{01} jsou počty dvojic zařazených v U do společného a ve V do rozdílného shluku, respektive naopak. Graficky je to naznačeno v obrázku 4.2.

Jedná se o symetrické metriky, tudíž vhodné pro naše použití, kdy nelze říci, co je správná odpověď. Pro výpočet vzájemné informace (anglicky *mutual information*) a Rand indexu využíváme implementace z knihovny *scikit-learn*[6], obě ve verzi *adjusted*, tedy takové, která zohledňuje vliv náhody v kontextu počtu shluků (např. neupravená vzájemná informace by pro výsledek analýzy s větším počtem shluků byla obecně vyšší nehledě na shodu porovnávaných verzí). Definice jsou následující:

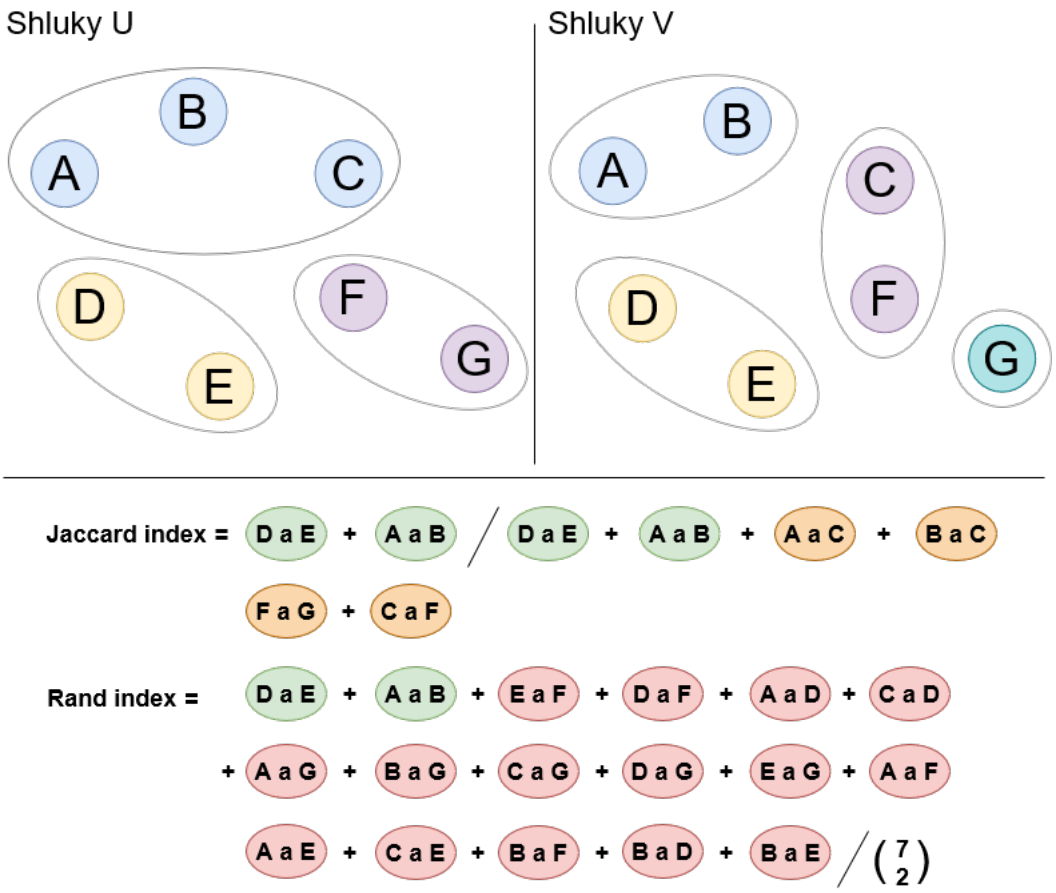
$$adjusted\_RI = (RI - \mathsf{E}\,(RI))/(max(RI) - \mathsf{E}\,(RI)),$$

kde RI je Rand Index a E vyjadřuje střední hodnotu, tedy máme hodnotu 1 pro kompletní shodu verzí shluků, hodnotu blížící se 0 pro náhodné rozložení a hodnotu jdoucí k -1 pro úplně rozdílné shluky.

$$adjusted\_MI(U,V) = \\ = [MI(U,V) - \mathsf{E}\,(MI(U,V))]/[avg(H(U),H(V)) - \mathsf{E}\,(MI(U,V))],$$

kde MI je skóre vzájemné informace, H je entropie, U a V jsou dvě verze shluků a $avg()$ vyjadřuje aritmetický průměr. Opět platí, že úplná shoda má hodnotu 1 a v případě náhodných uspořádání se pohybujeme kolem 0.

[6] `https://scikit-learn.org/stable/index.html`



Obrázek 4.2: Ukázka Jaccard a Rand indexu

4.2.2 Evaluace muzikologickou znalostí

Jak již bylo popsáno v sekci 4.2.1 chybí nám správné odpovědi k celkovému rozložení tradic či komunit. Pro alespoň částečnou „zkoušku příčetnosti“ (*sanity check*) můžeme využít existujících popisů některých dílčích fenoménů. Jde tedy o pozorování výsledků „globálního“ modelu se zaměřením na jeho chování na některých těch „lokálních“ částech a následnou konfrontaci s očekáváními.

Takové potenciální vodítko k hodnocení získaných výsledků nám poskytl muzikolog prof. David Eben z ÚHV FF UK. Vytipoval pro nás šest zinventarizovaných českých pramenů a nastínil jejich očekávané chování. Pro tři knihy používané klášterem sv. Jiří na Pražském hradě (signatury Cz-Pu XIV B 13, Cz-Pu VI E 4c, Cz-Pu VI G 11) se dá očekávat příklon k jihoněmeckým klášterním pramenům, mimo jiné díky benediktinské vazbě. A zároveň, i přes geografickou blízkost, by nebyly překvapivé rozdíly vůči vybraným liturgickým svazkům vzniklým pro pražskou katedrálu či v rámci pražské diecéze, jmenovitě Cz-Pn XII A 24, Cz-Pn XV A 10 a Cz-Pu XIV A 19. To může vycházet i z jisté autonomie či výsadního postavení tohoto kláštera, v době vzniku vybraných pramenů ještě přetrvávajícího[7].

To všechno jsou historické, kulturní či muzikologické ukazatele, které se snadno stanou příliš komplexními pro získání jasných odpovědí. Navrhovat takovéto evaluační postupy je proto náročné a vyžaduje to hlubokou znalost některého kusu repertoáru, navíc zasazenou do správných politicky a církevně historických kulis.

[7] Jak zmiňuje např. Pacovský (2023).

V tabulkách 4.1 a 4.2 vidíme míru překryvu jednotlivých pramenů, co do konkrétních zpěvů i do zastoupení svátků, jejichž repertoár je přítomen. Získáváme tak představu o tom, nakolik je relevantní z chování těchto knih něco usuzovat. Třicet sedm sdílených svátků by v případě jejich kompletního unikátního repertoáru přineslo 30 antifon a 12 až 15 responsorií. Průměrně šest kusů na den se může jevit jako překvapivě nízké číslo, ovšem zkušenost z knih ukazuje, že často je specifická pro daný den pouze část repertoáru (někdy jen připomínka jedním zpěvem) či jsou některé kusy ve dni zpívány opakovaně (například antifony v prvních a druhých nešporách či responsoria ve více nokturnech), což odhaluje i pohled na počet záznamů v knize vs. počet unikátních CID. Společné svátky jsou tvořeny repertoárovým základem temporálu – adventem a Vánocemi.

pramen	**počet CID**	**sdílená CID v rámci instituce**	**sdílená CID všech**
XII A 24	543	437	204
XIV A 19	853		
XV A 10	1298		
XIV B 13	1347	712	
VI E 4c	1667		
VI G 11	1842		

Tabulka 4.1: Míra sdílení zpěvů mezi šesti českými prameny

pramen	**počet svátků**	**sdílené svátky v rámci instituce**	**sdílené svátky všech**
XII A 24	55	44	37
XIV A 19	98		
XV A 10	195		
XIV B 13	127	104	
VI E 4c	151		
VI G 11	362		

Tabulka 4.2: Míra sdílení svátečního repertoáru mezi šesti českými prameny

4.3 Pokusy

Zdrojový kód reprodukující následující výsledky lze nalézt v jednotlivých souborech podle metody v elektronické příloze ve složce *research*.

4.3.1 Shluková analýza pomocí metody DBSCAN

V této sekci ukážeme chování zástupce *density-based* algoritmů pro shlukovou analýzu, DBSCANu, na grafu sestaveném ze všech 250 větších pramenů při použití Jaccardovy vzdálenosti pro výpočet rozdílů mezi prameny a také na menších grafech pro svátky sestavené stejným principem z knih, které k svátkům mají záznamy.

Zvolená implementace algoritmu DBSCAN z knihovny *scikit-learn* umožňuje výsledek ovlivnit skrze dva parametry, jak jsou popsány v sekci 4.1.1. Pro matici vzdáleností (bylo proto potřeba Jaccardovu vzdálenost „obrátit“ jejím odečtením od 1) algoritmus vrací seznam rozřazení do shluků, přičemž přiřazení hodnoty -1 znamená, že je daný vzorek (v našem případě pramen) považován za šum.

V první fázi pokusů jsme proto procházeli možné kombinace parametrů a sledovali, jaké je procento pramenů řazených mezi šum. Ze všech 54 zkoušených kombinací parametrů

- `eps`: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 a 0.9
- `min_sample`: 2, 3, 4, 5, 6 a 7

pouze pro dvojici 0.9 a 2 vyšel počet šumných vzorků menší než 125 (tedy méně než polovina z celkových 250 pramenů), konkrétně 112. Zároveň tato volba parametrů je ta, která klade nejmenší nároky na podobnost pramenů (vzdálenost vrcholů do 0.9 je považována za sousedskou) a zároveň nejsnáze konstruuje shluk (na jeho založení stačí dva blízké prameny).

Je tedy zjevné, že pro výpočty na všech datech s použitím Jaccardovy vzdálenosti odečtené od 1, je algoritmus nevhodný. Matice vzdálenosti obsahuje obecně vysoká čísla (blízká 1) a všechna velice podobných hodnot, což není vhodná situace vzhledem k popsanému fungování algoritmu, který hledá centra s vyšší hustotou.

Nevhodnost na velkém kusu dat ovšem nemusí nutně implikovat nepoužitelnost pro menší kusy repertoáru. Provedli jsme proto podobné měření s `eps`=0.9 a `min_sample`=2 separátně pro každý svátek s alespoň 10 záznamy v naší sadě z antifon a responsorií (takových svátků je 918). Z těchto 918 má méně jak polovinu pramenů v odpovídajícím grafu klasifikovanou jako šum pouze 48 svátků (jedná se z velké většiny o kusy z temporálu). To už ukazuje na celkovou nevhodnost zvoleného přístupu: pokud tradice existují, DBSCAN je nedokáže najít.

4.3.2 Louvein detekce komunit

V této sekci popíšeme experimenty prováděné s Louvein (lovaňským) algoritmem (viz 4.1.2). Nejdříve bylo třeba určit jeho vhodnou parametrizaci, následně jsme prováděli výpočty na větším i menším množství dat a zkoumali, nakolik jsou v jednotlivých případech vrácená řešení stabilní. Na závěr jsme se podívali specificky na sváteční repertoár sdílený napříč šesti vytipovanými českými prameny a pozorovali jejich výsledné rozložení v nalezených komunitách.

Hledání hodnoty parametru rozlišení

V první fázi pokusů s Louvein algoritmem bylo třeba najít jeho vhodnou parametrizaci, což v tomto případě zahrnovalo parametr jediný – míru rozlišení (`resolution`). Vyšší hodnota než 1.0 podporuje vznik většího počtu komunit, nižší počtu menšího. Výpočet probíhal na síti, jejíž vrcholy byly jednotlivé prameny (celkem 250) a váha hran mezi nimi byla vypočítána jako Jaccardova vzdálenost mezi množinami zpěvů v nich zapsaných. Pro každou hodnotu rozlišení jsme nechali sestavit padesát komunitních variant a na nich počítali stabilitu. Její vývoj, včetně změn v průměrném počtu nalezených komunit, lze pozorovat v grafu 4.3.

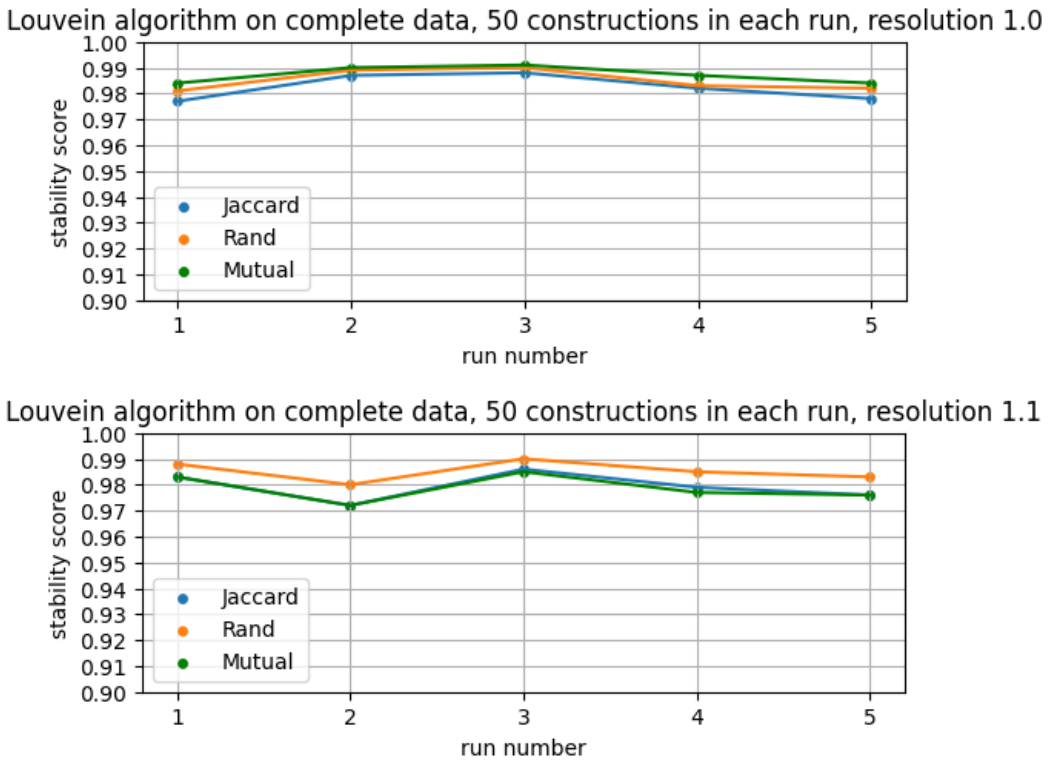
Obrázek 4.3: Graf hodnot zvolených metrik při změně parametru rozlišení (osa y nezačíná v 0)

Síť ze všech dat

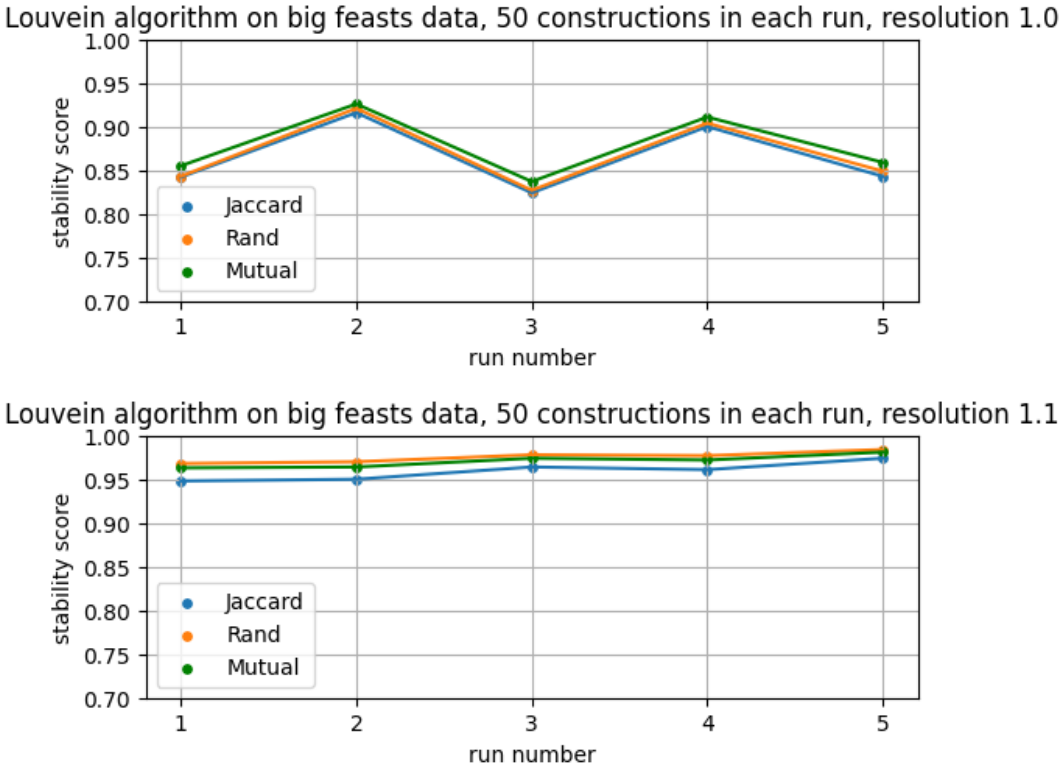
Následně jsme provedli měření stability nalezených komunit pomocí tří vybraných metrik při konstrukci komunit na síti ze všech dat (250 pramenů), s Jaccardovou vzdáleností pro váhy hran a s rozlišeními 1.0 a 1.1 (hodnoty byly zvoleny na základě předchozího měření, jehož výsledek zobrazuje graf 4.3). Stabilitu sestavených komunit ukazuje graf 4.4. V něm zároveň vidíme vliv náhody v běhu algoritmu. Prvotní náhodné určení center má na výslednou podobu komunit vliv – proto také došlo vždy k padesáti sestavením komunit na síti s různým náhodným semínkem (anglicky *random seed*) a následnému porovnání stability mezi nimi. Průměrný počet komunit pro hodnotu 1.0 je čtyři a pro 1.1 šest.

Síť pro velké svátky

Pokračovali jsme s výpočty pro deset největších (nejčetnějších) svátků (viz přehled 3.2) Vybrali jsme všechny prameny, které obsahují záznam ke každému z těchto deseti svátků, takových bylo 51, a postavili z nich síť s pomocí zpěvů k těmto svátků (1293 CID), hrany jsou opět váženy Jaccardovou vzdáleností. Nyní jsme aplikovali stejný postup jako u sítě ze všech dat – rozlišení 1.0 a 1.1, pro obě hodnoty pět kol, každé po padesáti běžích. Výsledná stabilita je vidět v grafu 4.5, průměrné počty komunit jsou tři a devět.



Obrázek 4.4: Graf hodnot metrik stability v síti všech dat v 5 různých pokusech (osa y nezačíná v 0)

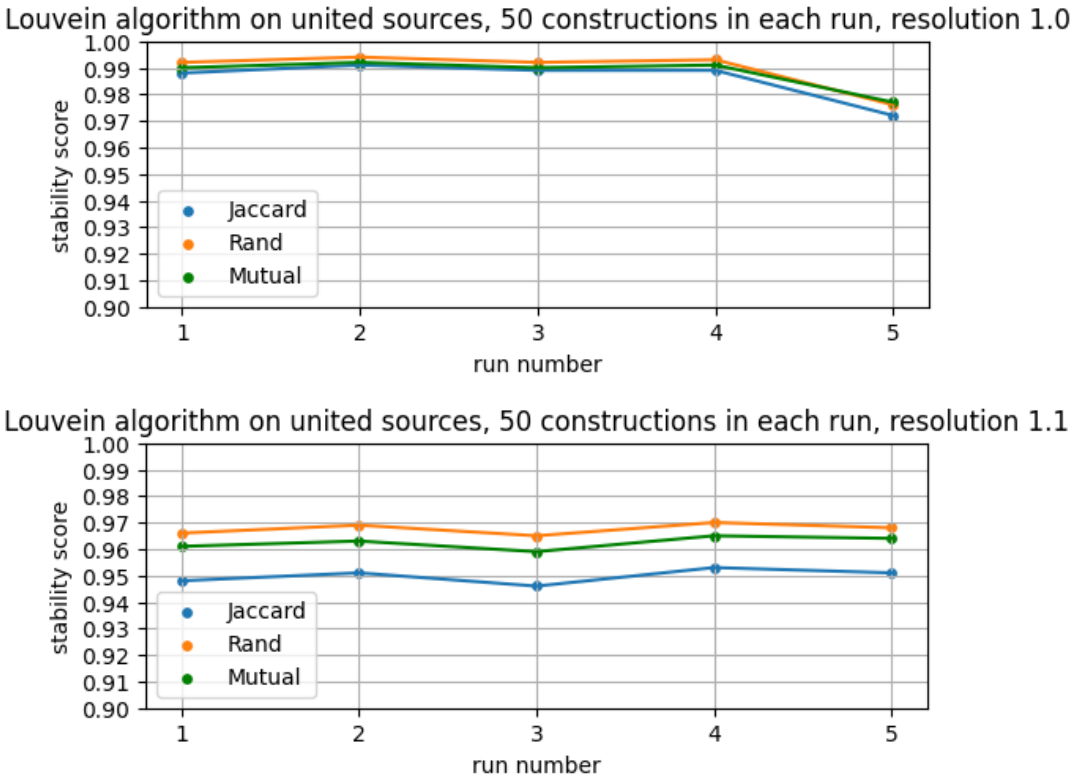


Obrázek 4.5: Graf hodnot metrik stability v síti z dat k 10 největších svátcích v 5 různých pokusech (osa y nezačíná v 0)

Slučování vrcholů

Sloučíme-li obsah liturgických knih z jednoho místa, ze stejného století a stejného institucionálního původu, známe-li všechny tyto vlastnosti, vždy do jednoho vrcholu sítě (tedy logikou do jednoho fiktivního agregovaného pramene), pak zredukujeme počet vrcholů z 250 na 193. Cílem je se dívat víc na místa v čase než na konkrétní svazky, které mohou geografickou interpretaci rozbíjet (např. má-li klášter zimní a letní část sanktorálu zapsanou ve dvou svazcích, protože by se do jedné knihy z praktických důvodů množství repertoár nevešel).

Na takto vytvořené síti můžeme provést měření použité v předchozích dvou sekcích. Volba rozlišení 1.0 a 1.1 je opět dána výsledky hledání parametru zobrazenými v grafu 4.3. Tentokrát je průměrný počet vrácených komunit pro 1.0 čtyři a pro 1.1 devět. Hodnoty skóre stability ukazuje graf 4.6.



Obrázek 4.6: Graf hodnot metrik stability v síti se sloučenými vrcholy na základě shodných vlastností v 5 různých pokusech (osa y nezačíná v 0)

Sítě okolo českých pramenů

Posledním krokem byl pohled na šest českých pramenů (Cz-Pn XII A 24, Cz-Pn XV A 10, Cz-Pu XIV A 19 a Cz-Pu XIV B 13, Cz-Pu VI E 4c, Cz-Pu VI G 11), jak jsou popsány podrobněji v sekci 4.2.2. Sestavili jsme síť pouze z takových liturgických knih, které obsahují zpěvy k některému ze 37 sdílených svátků a do výpočtu hran pomocí Jaccardovy vzdálenosti jsme zahrnuli pouze zpěvy k těmto svátkům. Zůstalo nám 153 pramenů (je pravděpodobné, že jsme z celkových 250 pramenů vyřadili sanktorály, neboť sdílené svátky byly adventní a vánoční). Z tohoto postupu se vytrácí kontrola stability. Dovolujeme si tento

krok provést, neboť míra jistoty, s jakou na zkoumaném materiálu (celém i jeho částech) Louvein algoritmus odpovídá, je vysoká. Rozebereme zde podobu výsledku pro různé hodnoty parametru rozlišení.

- hodnoty 0.2, 0.3, 0.4, 0.5 a 0.6
  - čtyři komunity
  - jedna veliká o 150 zpěvech a 3 po jednom prameni – *E-E, L. III. 3.*, *E-SAu Ms 2637* a *D-W Cod Guelf Helmst 1008*
- hodnota 0.7
  - pět komunit
  - největší po 145 pramenech, dále jedna po pěti (*B-Br Ms IV/473*, *E-BAR Cj. 95*, *E-Mn Mss/1361*, *F-Pn Lat. 909*, *F-Pn NAL 01235*) a zbylé tři po jednom prameni viz nižší hodnoty rozlišení
- hodnota 0.8
  - pět komunit
  - největší po 140 pramenech, tři jednoprvkové viz výše a deset pramenů (tři španělské, šest francouzských a jedna belgická signatura)
- hodnota 0.9
  - pět komunit
  - tři jednoprvkové komunity stejného složení, dále 113 pramenů včetně všech českých a skupina 37 pramenů rozmanitého původu
- 1.0
  - sedm komunit
  - komunita o 65 pramenech obsahující *XII A 24* a prameny ze všech koutů Evropy
  - komunita o 47 pramenech španělského, francouzského a italského rázu
  - komunita o 30 pramenech obsahující zbylých pět českých pramenů a dále hlavně německé a rakouské prameny
  - tři jednoprvkové viz výše a dále komunita osmi pramenů s rozmanitým původem
- hodnota 1.1
  - dvacet komunit
  - již velké roztříštění, největší komunity mají 45, 22 a 10 pramenů
  - dochází k rozdělení českých pramenů
    - *Pn XV A 10* a *Pu XIV A 19* spolu s *D-W 29 Helmst.*, *GB-Ob MS. Canon. Liturg. 202*, *MA Impr. 1537*, *PL-WRu R 503*, *SK-BRsa SNA 2* a *TR-Itks 42*

    - svatojiřské *Pu VI G 11*, *Pu VI.E.4c* a *Pu XIV B 13* jsou v komunitě společně s dvěma německými, třemi rakouskými a jedním švýcarským pramenem
    - *XII A 24* je v komunitě pouze s *D-MZb A* a *PL-Kkar 2 (Rkp 14)*

- vyšší hodnoty parametru rozlišení
  - vysoká roztříštěnost, vyčleňování velkého množství jednoprvkových komunit
  - až po hodnotu 1.8 zůstává v každé variantě komunitního rozložení jedna největší skupina průměrně 40 pramenů tvořená hlavně španělskými a francouzskými prameny, ovšem s příspěvky dalších, nikoliv českých, pramenů
  - šest vytipovaných českých pramenů nacházíme v komunitách nanejvýš po dvou a to vždy podle instituce

Dále jsme sestavili síť pouze z pramenů, které obsahují alespoň 30 ze sdílených 37 svátků. To z důvodu snahy o odstínění vzdálenosti mezi některými prameny způsobené kvantitou uvažovaných svátků v prameni. (Je rozdíl, je-li v knize 30 ze 37 svátků a k nim odpovídající počet zpěvů či vyskytuje-li se pouze repertoár ke dvěma svátkům. Při výpočtu váhy hrany mezi takovými knihami to vytváří v Jaccardově vzdálenosti vysokou hodnotu ve jmenovateli, která stíní i případnou shodu v těch sdílených dvou svátcích.) Získali jsme tak síť o 81 vrcholech, kde pracujeme s 1071 různými CID. Následující přehled popisuje chování výstupu algoritmu vzhledem k hodnotám parametru rozlišení.

- hodnoty 0.2, 0.3, 0.4, 0.5 a 0.6, 0.7, 0.8, 0.9
  - jedna komunita
- hodnota 1.0
  - tři komunity
  - 39, 34 a 8 pramenů
  - všech šest českých se drží spolu ve skupině s převážně sousedními zeměmi a dále několika dalšími prameny ze zemí vyjma Španělska
- hodnota 1.1
  - dvacet komunit
  - svatojiřské *Pu VI.E.4c* a *Pu XIV B 13* jsou v komunitě společně s dvěma německými, třemi rakouskými a jedním švýcarským pramenem
  - zbývající klášterní *Pu VI G 11* je spolu s dalšími sedmi, kde máme dva španělské a dva francouzské prameny
  - katedrální *Pn XV A 10*, *Pu XIV A 19* a *XII A 24* se chovají stejně jako na předchozí síti při hodnotě 1.1

- vyšší hodnoty
  - opět vysoké roztříštění a jednotlivé prameny ve vlastních komunitách
  - zajímavá je skupina osmi pramenů včetně *VI G 11* popsaná už u hodnoty 1.1, která se při zvyšování rozlišení netrhá
  - rozložení neobsahují rozporný výskyt našich šesti českých pramenů

Shrnutí pokusů s algoritmem Louvein

Z pohledu stability se Louvein algoritmus při parametru rozlišení 1.0 a 1.1 jeví slibně, hodnoty všech tří skóre jsou nad 0.8 a ve většině případů i nad 0.9. Je ovšem potřeba mít na paměti, že toto nic nevypovídá o vlastní kvalitě odpovědí, ale pouze o jistotě algoritmu. Drobnou kvalitativní výpovědní hodnotu mají výsledky části s českými prameny. Zde nedochází k rozporu v jejich vzájemném uspořádání, ovšem některé nalezené skupinky by bylo, pro větší jistotu či nějaký závěr, třeba detailněji prozkoumat. (To bohužel vyžaduje muzikologické znalosti už nad rámec rozsahu této práce, nicméně poskytujeme pro takovou analýzu příslušné nástroje – viz kap. 5).

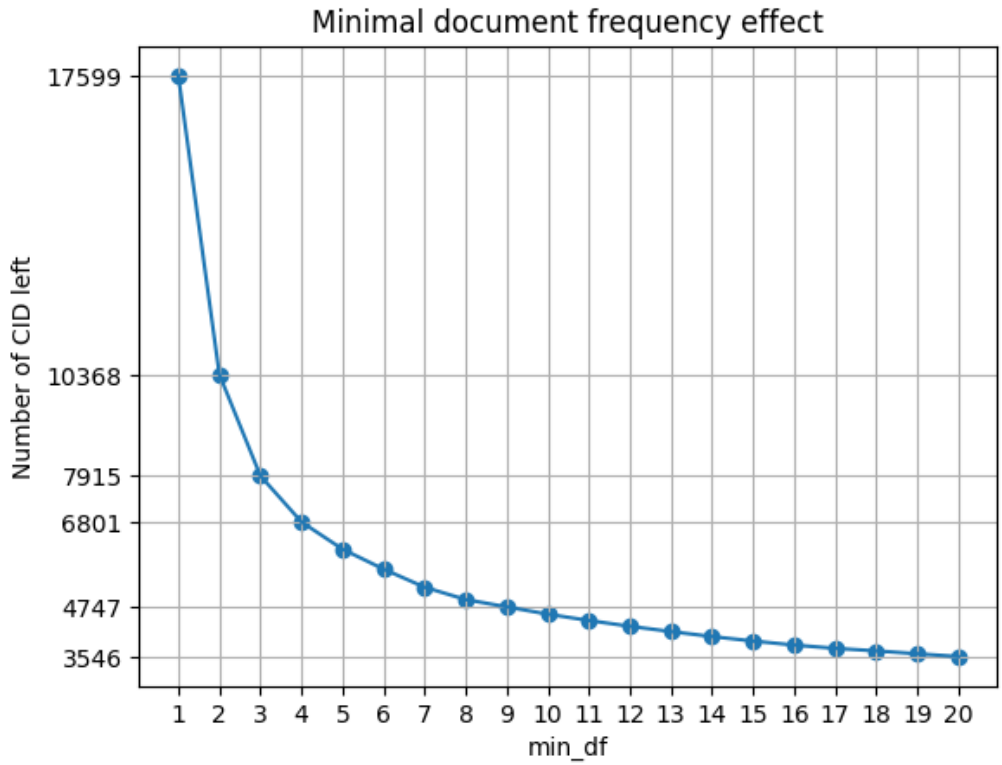
4.3.3 Modelování témat na pramenech

V následujících řádcích bude cílem představit pokusy prováděné s pomocí pravděpodobnostního modelu *Latent Dirichlet Allocation* používaného k modelování témat v korpusu. V první fázi probíhalo hledání vhodné parametrizace v podobě volby nastavení „přísnosti“ předpočítané matice výskytu slov v dokumentech a také maximálního počtu iterací výpočtu. Následně jsme nechali tematické rozdělení napočítat na různém počtu témat a pozorovali stabilitu nalezených řešení, a to jak na všech dostupných datech, tak také pouze na datech z deseti nejrozšířenějších svátků. Dále nás zajímala podoba výsledků pro dvě témata vzhledem k monastickému resp. sekulárnímu původu pramenů. Na závěr jsme se pokusili sjednotit prameny stejného místa, času a cursu do jednoho dokumentu (viz sekce 4.3.2) a sledovali stabilitu tohoto přístupu.

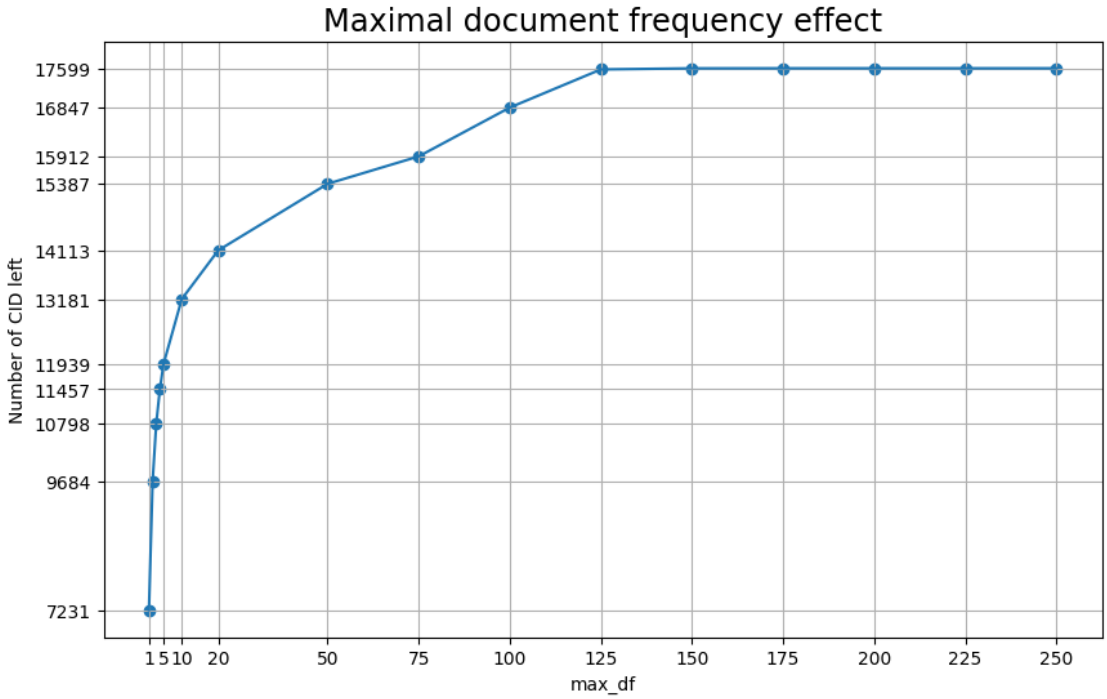
Parametry pro modelování témat

Zvolená implementace pro výpočet matice *prameny* × *Cantus ID* nám umožňuje ovlivnit započítaná slova (tedy pro nás právě Cantus ID) pomocí dvou parametrů – `max_df`, udávající maximální počet dokumentů (pro nás pramenů), které obsahují dané slovo, aby bylo do matice uvažováno, a `min_df`, který naopak říká, jaký je minimální počet dokumentů, ve kterých se slovo vyskytuje, aby bylo takové započítáno. Vliv těchto dvou parametrů na počet Cantus ID je zobrazený v grafech 4.7 a 4.8.
Na základě těchto pozorování použijeme tři modely *CountVectorizer* (viz 4.1.3) pro předpočet matice:

- `all_count_vec` – všechna slova (17 599 CID)
- `less_count_vec` – slova v alespoň dvou dokumentech (10 368 CID)
- `smallest_count_vec` – slova v alespoň osmi dokumentech (4 924 CID)

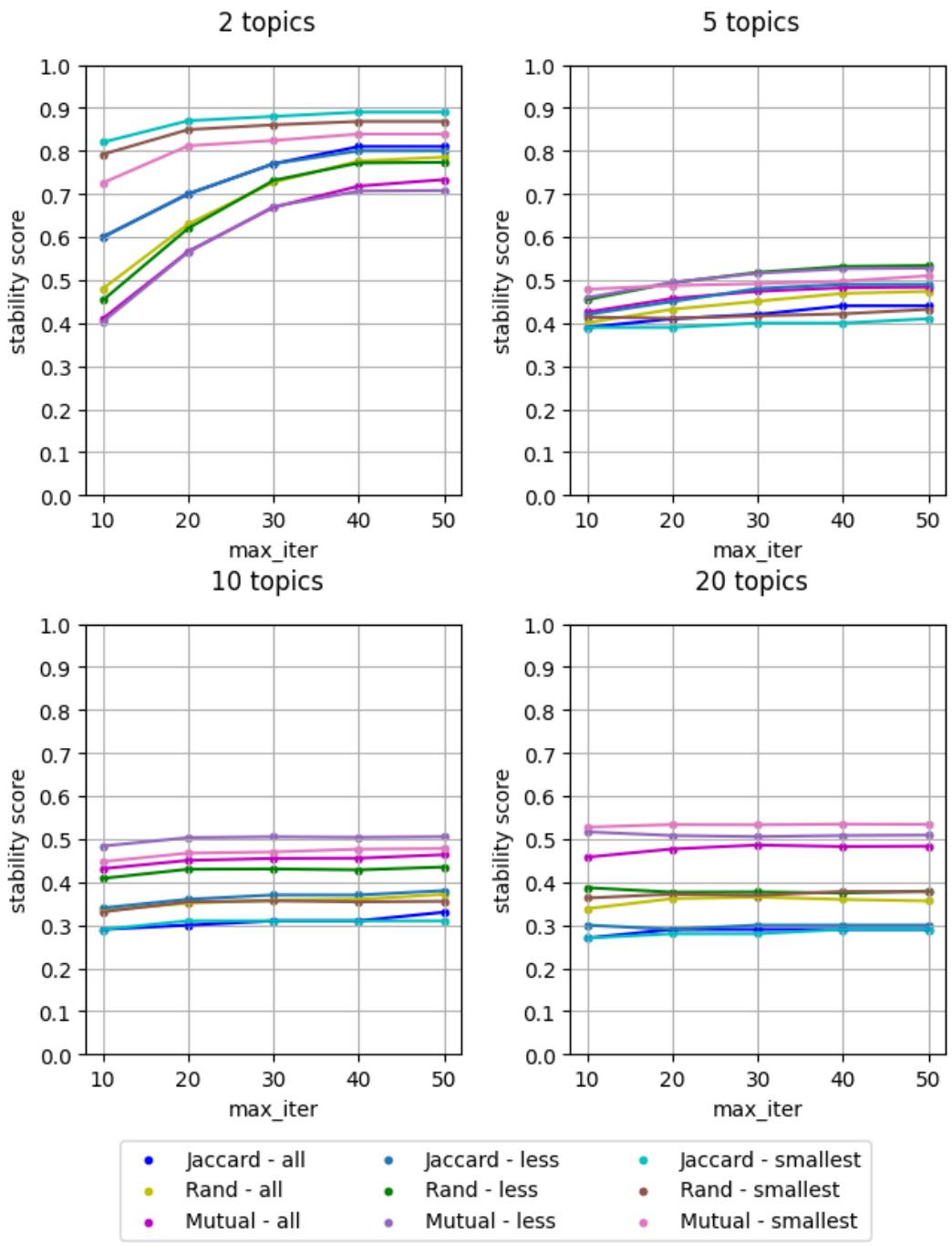


Obrázek 4.7: Graf vlivu minimální frekvence dokumentů na počet Cantus ID, když `max_df` je 250



Obrázek 4.8: Graf vlivu maximální frekvence dokumentů na počet Cantus ID, když *min_df* je 0.0

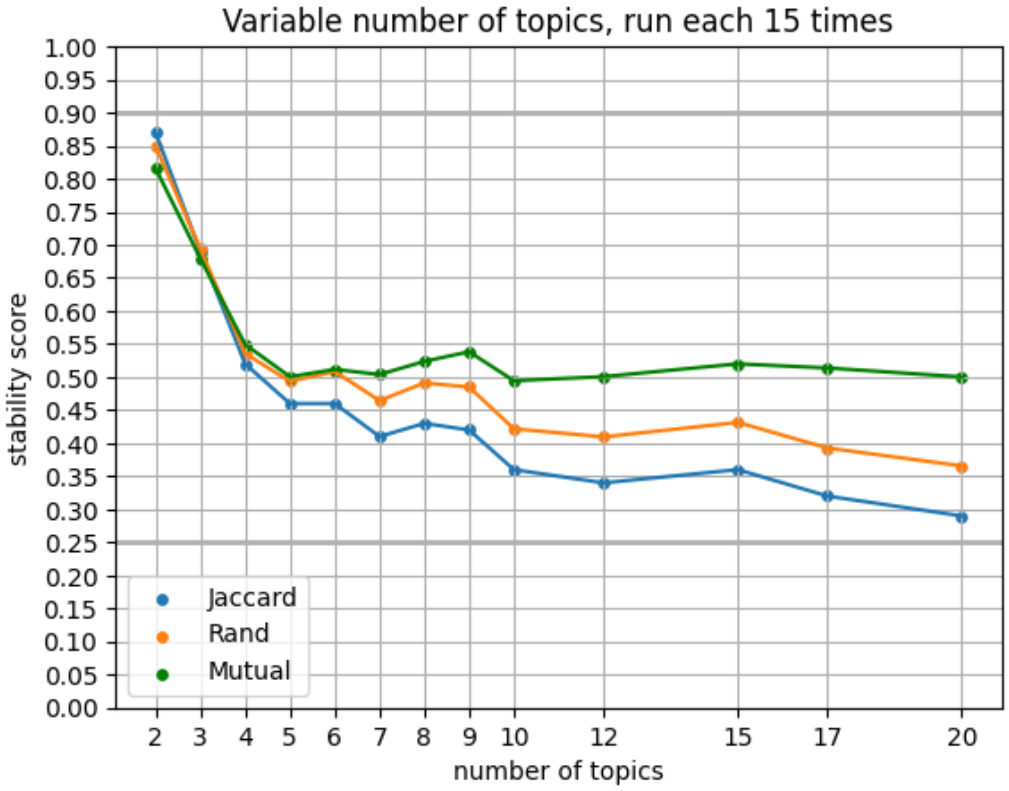
Tím jsme si ujasnili varianty jednoho ze dvou parametrů, jejichž vliv na stabilitu hledaných komunit jsme v dalším kroku zkoumali. Druhým parametrem byl maximální počet iterací. Následující grafy 4.9 ukazují, jak se měnila skóre (v grafu rozlišeno barevně) jednotlivých metrik stability představených v sekci 4.2 v závislosti na volbě počtu iterací (osa x) a modelu pro matici frekvence slov v dokumentech (odlišeno barvou). Pro každou konfiguraci jsme sestavili (a následně porovnali) 10 modelů.



Obrázek 4.9: Grafy vlivu parametrů na skóre stability pro 2, 5, 10 a 20 témat

Dokumentová struktura pro všechna data

Z přehledu 4.9 vidíme vliv nastavení modelu *CountVectorizer*. To nám poskytuje představu o tom, pro který počet témat použít který ze zkoumaných modelů. Pro parametr `max_iter` vyšla nejstabilněji hodnota 40, a to nehledě na poptávaný počet témat. V rámci volby předpočtu matice jsme použili pouze dvě varianty – `smallest_count_vec` pro 2, 3 a 4 témata a `less_count_vec` pro zbylé počty. Výsledný vliv těchto parametrů na stabilitu modelu LDA ukazuje graf 4.10, pro každý počet témat jsme model sestavili a natrénovali patnáctkrát.



Obrázek 4.10: Graf skóre stability pro větší množství témat na všech datech

Dokumenty pro deset největších svátků

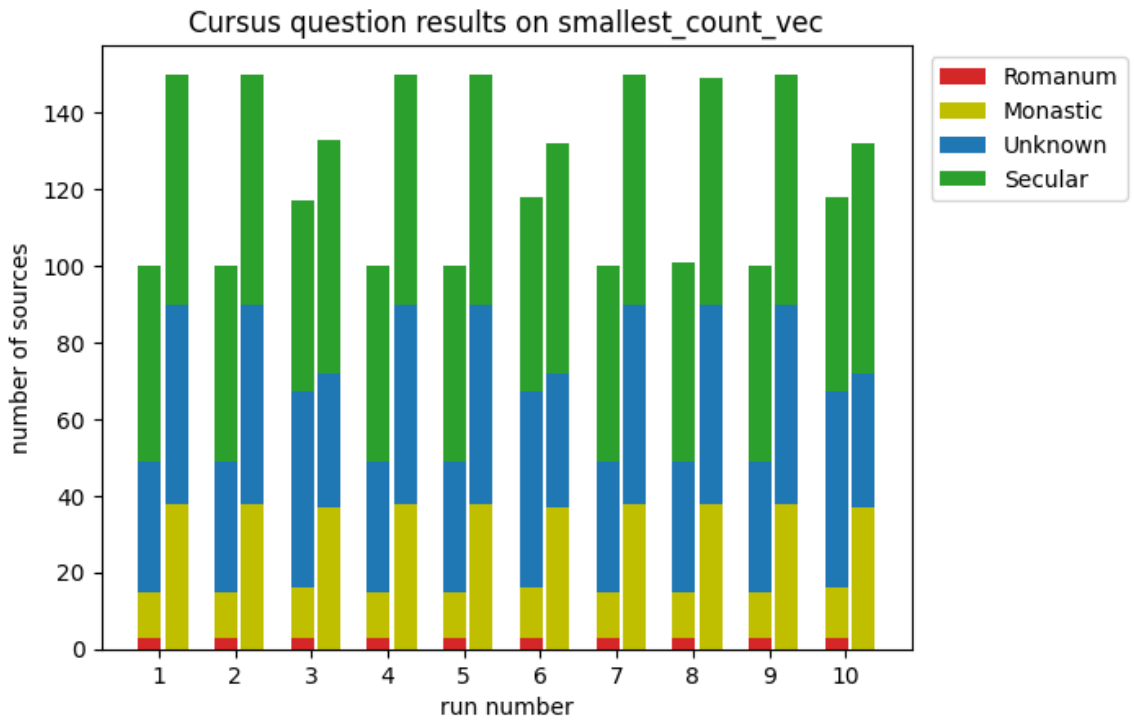
Dalším krokem bylo omezení dat ve výpočtu na ta spojená s deseti největšími (co do počtu záznamů) svátky (k nahlédnutí v diagramu 3.2). Tím jsme snížili počet dokumentů v kolekci z 250 na 51 a počet slov (Cantus ID) z celkových 17 599 na 1 293. Bylo proto nutné vytvořit novou matici s frekvencemi slov v sníženém počtu dokumentů, do které jsme započítali všechna slova (tedy zpěvy) příslušná daným pramenům pro vybrané svátky. Sestavili a natrénovali jsme 15 LDA modelů, z každého vzali distribuci témat pro dokumenty (knihy) a knihu přiřadili k tématu, které mělo nejvyšší pravděpodobnost v získané distribuci. Výsledek je patrný z grafu 4.11.

Obrázek 4.11: Graf skóre stability pro množství témat na datech k 10 největším svátkům

Dvě témata podrobněji

V případě zkoumání výsledku dvou témat se nabízelo podívat, i přes částečně chybějící data (ne každý pramen má tento údaj uveden), na rozložení institucionálního původu pramenů napříč nalezenými skupinami. Dělení na monastické a sekulární prameny by bylo nadějnou interpretací výsledků, neboť se jedná o dvě poměrně dobře definované kategorie s odlišnými liturgickými potřebami (kláštery obecně pro jeden den potřebují zpěvů více). Zvolili jsme výpočet nad všemi prameny a všemi svátky, neboť se jeví stabilněji než repertoár 10 největších svátků. Sestavili jsme 10 variant (neboť 2 témata se ukazují jako ještě poměrně stabilní) komunit na stabilitou nejlépe vycházejícím nastavení zachovávání slov v k výpočtu používané matici – `smallest_count_vec` (viz 4.3.3).

Distribuci jednotlivých možností pro tuto základní institucionální příslušnost (sekulární vs. monastický původ, položka *cursus*) pramenů v rámci deseti běhů ukazuje následující diagram 4.12. V něm vidíme v zásadě dvě nalezené verze rozložení (jednu v pokusech 3, 6 a 10 a druhou ve zbylých), kde ovšem ani jedna neukazuje na nějaké monasticko-sekulární rozložení. České prameny se v obou variantách chovají stejně, VI G 11 se odděluje od zbývajících pěti. Vzhledem k tomu, že počítáme na všech zpěvech, toto může být dáno prostě jenom tím, které svátky jsou v VI G 11 zapsány mimo 37 kompletně sdílených. Z tabulky 4.2 vidíme, že právě VI G 11 je co do počtu zapsaných svátků nejbohatší.



Obrázek 4.12: Rozložení vlastnosti cursus v komunitách pro dvě témata

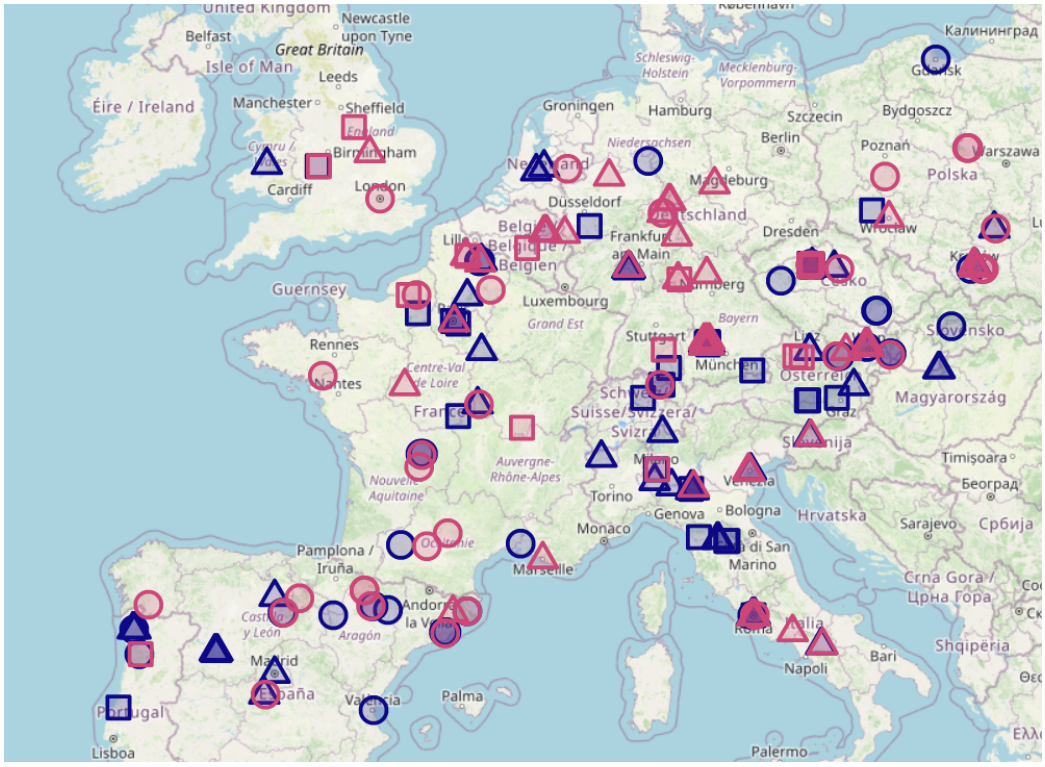
Následující mapa 4.13 ukazuje, byť ne pro všechny prameny známe lokalitu jejich původu, že rozdělení do dvou komunit, navržené tematickým modelováním, nebude v tomto případě mít ani jasnou podporu geografie. Různé tvary bodů označující jednotlivé prameny nesou informaci o původu – trojúhelník pro sekulární prameny, čtverec pro monastické a kolečko pro ty s původem neznámým.

Sloučení dokumentů

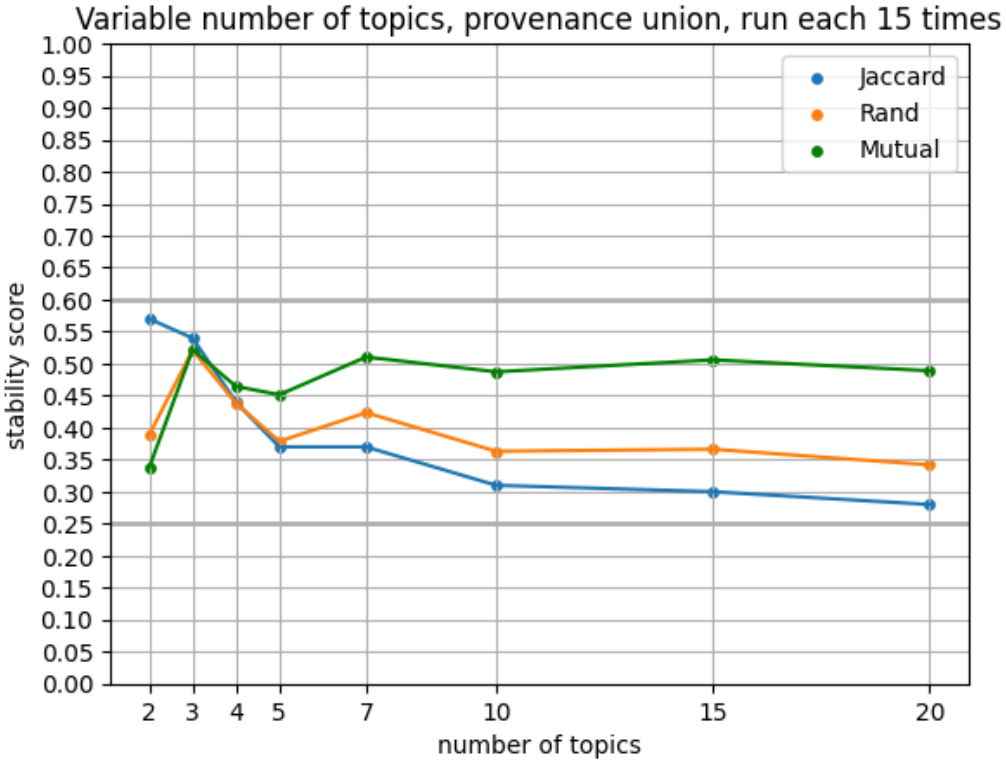
Poslední zkoušená myšlenka pramenila z úvahy o sjednocení knih z jednoho místa, století a z klášterního/farního prostředí do jednoho dokumentu. To by mohlo přestat věci geograficky zbytečně tříštit a řešit i problémy, jakými v našich úvahách mohou být např. zimní a letní části sanktorálů (viz také 4.3.2).

Vytvořili jsme proto novou kolekci dokumentů a to tak, že knihy, jejichž století vzniku, identifikátory porvenance a cursus, které nebyly neznámé, se shodovaly, byly vnímány jako jeden dokument. Jakmile byla některá ze zmiňovaných tří položek neznámou, pramen se vyčlenil jako samostatný dokument. Tím jsme získali 193 dokumentů. Stabilitu nalezených komunit pro takto poslepované některé prameny ukazuje graf 4.14. Každá konfigurace byla spuštěna patnáctkrát a to s uvažováním těch slov (Cantus ID), která se vyskytují v alespoň dvou dokumentech (to v tomto případě konstrukce dokumentů bylo 9 915 ID).

V porovnáním s přístupem, že dokument je jedna liturgická kniha, je zřetelný zejména propad stability pro dvě témata, ale celkově jsou hodnoty nižší. Pro většinu modelů se pohybujeme v části škály, kde lze mluvit spíše o náhodě než o libovolně systematickém uspořádání.



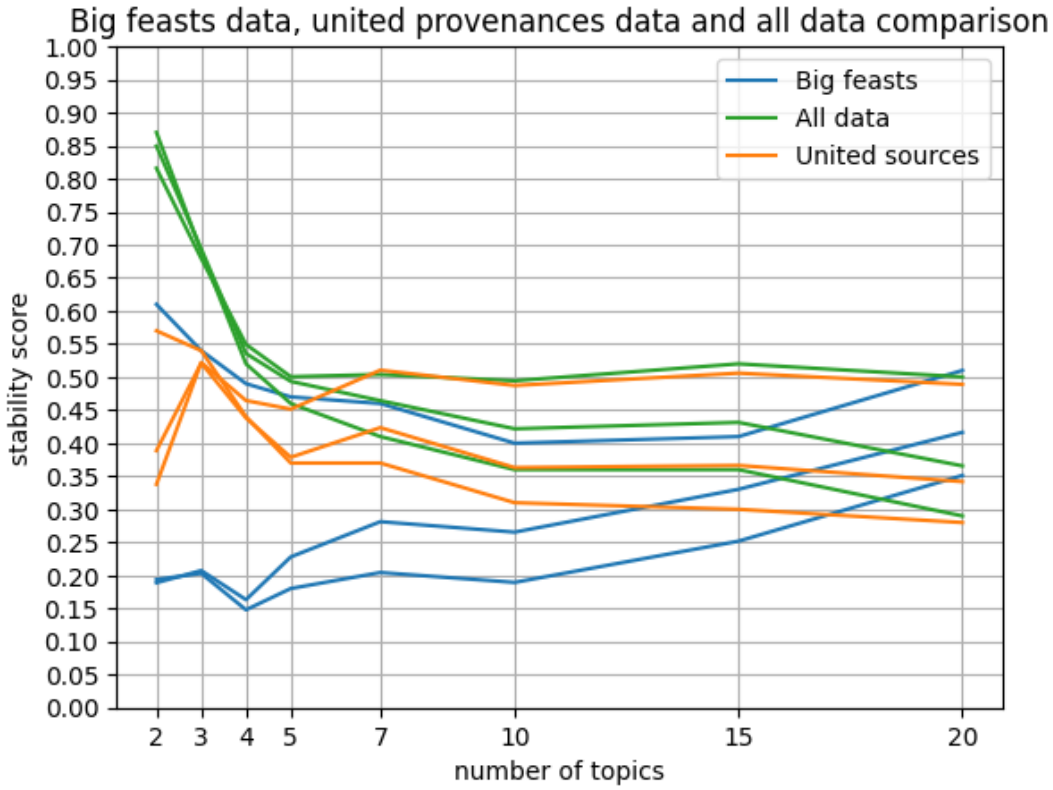
Obrázek 4.13: Mapa pramenů po namodelování dvou témat s barevně odlišenými komunitami.



Obrázek 4.14: Graf stability u dokumentů sjednocujících provenance pro jednotlivé počty témat

Shrnutí pokusů s tematickým modelováním

Graf 4.15 ukazuje porovnání mezi stabilitou nalezených skupin na všech datech, na datech k deseti největším svátkům a na těchto sloučených dokumentech. Vidíme, že výsledky představených tří přístupů ke kompletaci dokumentů pro LDA, se, mimo dvě témata pro všechna data, pohybují v části škály, která již vykazuje notnou náhodnost rozložení.



Obrázek 4.15: Graf skóre stability pro množství témat na datech k 10 největším svátkům, ke všem datům a k sjednoceným dokumentům

Vydat se směrem k pokusu na materiálu obsaženém v šesti vytipovaných českých pramenech (viz 4.2.2), který jsme provedli s algoritmem Louvein (viz 4.3.2), vzhledem k neuspokojivým výsledkům v otázce stability, obzvlášť při vyšším počtu témat, nelze považovat za relevantní postup. Nabízelo by se např. zkoumat od jakého počtu témat se české prameny vyskytují v různých komunitách a jakým způsobem, ovšem zkoumané výstupy by byly náhodným produktem, který by se při změně náhodného semínka vzápětí proměnil, proto jsme k němu nepřistoupili.

Celkově je ovšem neuspokojivé chování LDA docela překvapivé. Očekávali jsme, že potenciál rozklíčování i komplikovanějších situací, kde není každý dokument jenom z jedné tradice, se více projeví. Stav, kdy se maxima v distribuci měnila s každým sestavením modelu (a proto byly výsledky nestabilní), může naznačovat neexistenci předpokládaných tradic stejně tak jako nedostatek dat či nevhodnost jejich použití.

4.4 Existence průřezových tradic

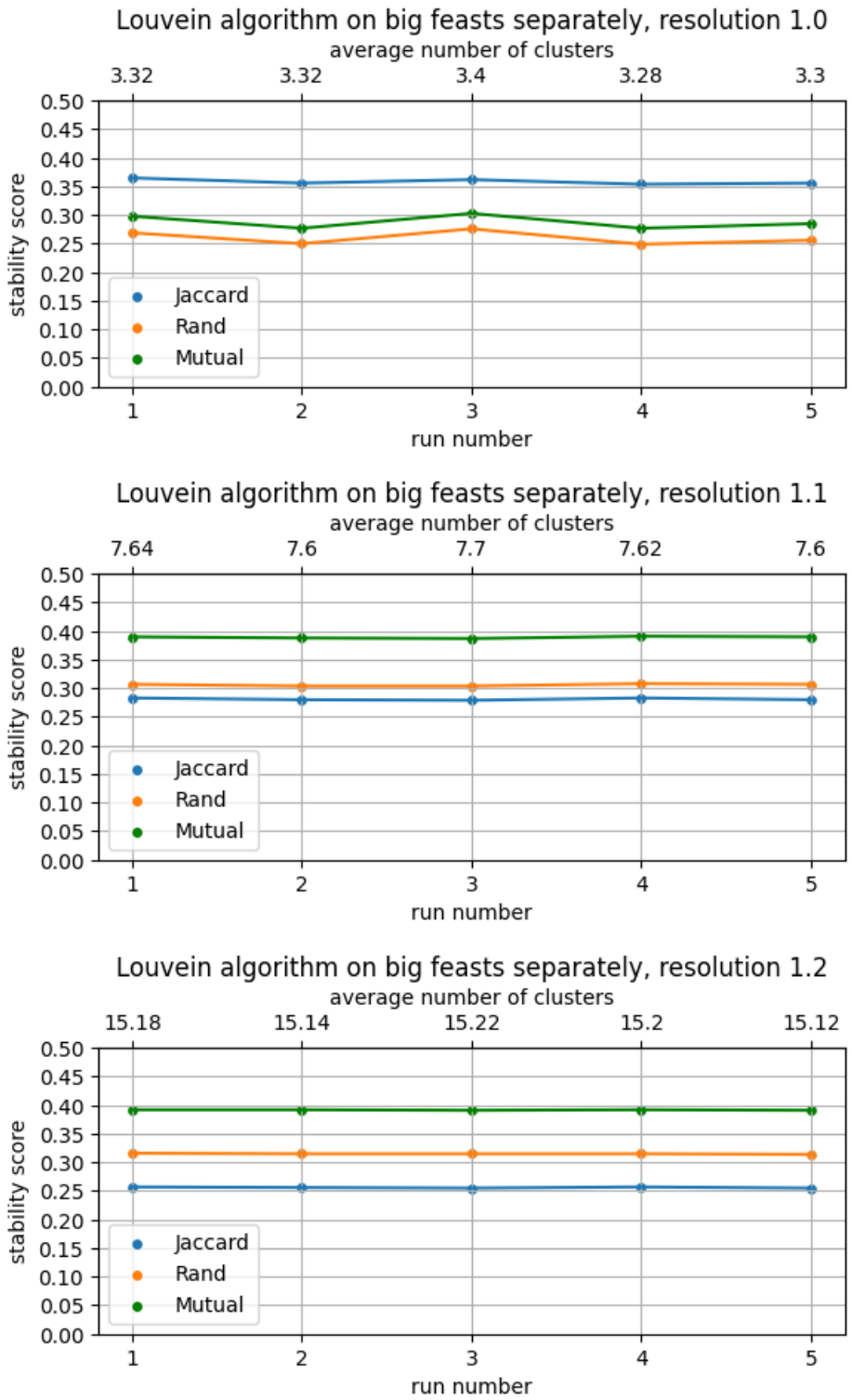
Jelikož výsledky pohledu na větší kusy materiálu (viz pokusy popsané v sekcích 4.3.2 a 4.3.3) nás úplně neuspokojily a rozhodli jsme se porovnat mezi sebou komunity nalezené po jednotlivých svátcích. Je totiž možné, že tradice popisované v muzikologické literatuře existují jen na úrovni jednotlivých svátků a ne nutně celých pramenů.

Pro tyto pokusy použijeme deset největších svátků (viz 3.2) a 51 pramenů, které obsahují repertoár ke všem z nich. Jsou vhodným vzorkem pro celý gregoriánský chorál. Jedná se o největší zástupce co do počtu záznamů, ale také o zástupce těch nejvýznamnějších svátků, jakými jsou v křesťanském světě Vánoce a Velikonoce, doplněné o další základní prvky, kterými jsou společná (univerzální) oficia mučedníků a apoštolů, a důležité světce: Pannu Marii a Jana Křtitele. Ze zkoumaných metod jsme na tento pokus použili algoritmus Louvein, neboť jeho výsledky jsou nejpřesvědčivější (nejvyšší stabilita a chování na českých pramenech podle muzikologických předpokladů). Navíc jako jediný už pro příbuzná chorální data přinesl výsledky: příspěvek Hileyho v knize *Embellishing the Liturgy: Tropes and Polyphony* (Planchart (2009)) k jehož výsledkům našli interpretaci Eipert s Mossem (Eipert a Moss, 2023a).

V první fázi jsme zkontrolovali, že výsledky žádného z deseti uvažovaných svátků se samy o sobě (komunitní rozložení porovnávaná pouze mezi sebou v rámci toho jednoho svátku) nechovají náhodně. Všechna tři skóre stability se ve všech případech pohybovala v rozmezí 0.8 až 1.0 pro zkoumané hodnoty parametru rozlišení 0.9, 1.0, 1.1 a 1.2, přičemž pro rozlišení 1.1 a 1.2 byl průměr přes všechny svátky pro všechny tři metriky nad hodnotou 0.9 a rozložení pramenů v komunitách nevykazuje žádné výrazné nerovnoměrnosti. (Podobně dobrá se ukázala i stabilita mnoha dalších jednotlivých svátků.)

Podpořeni těmito výsledky, jsme měřili, jak stabilita (nyní ve smyslu shody) klesne při porovnání komunit nalezených na různých svátcích. Sestavili jsme pět variant detekce komunit pro každý svátek. Získali jsme tak 50 variant komunit na 51 pramenech, což odpovídá pokusu 4.3.2, a ty jsme porovnali metrikami stability. Pro každou uvažovanou variantu rozlišení jsme provedli pět běhů (opět jako v pokusu 4.3.2). Výsledek tohoto porovnávání komunit nalezených v jednotlivých svátcích jsou zobrazeny v následujících grafech 4.16.

Naměřená skóre naznačují vysokou nejednotnost mezi jednotlivými velkými svátky. Tuto podmnožinu z dostupných dat, jak jsme popsali výše v této sekci, lze zároveň považovat za reprezentativní: pokud by signifikantní tradice po celých pramenech existovaly, potom by se měly odrazit právě v těchto svátcích, neboť tyto reprezentují výběr hlavních událostí liturgického roku. Z těchto dvou skutečností můžeme usuzovat, že repertoárové tradice nejsou na úrovni pramenů (prameny se pro každý ze svátků v našem měření chovaly jinak) a že zaměřit se na menší části je lepší přístup.



Obrázek 4.16: Graf stability porovnávání komunit jednotlivých velkých svátků (osa y nekončí v 1)

4.5 Diskuze

Z představených experimentů se jako nejslibnější jednoznačně jeví Louvein algoritmus. Jeho chování z pohledu stability je nejjistější. Přístup s pomocí modelování témat vykazuje při všech námi zvolených konstrukcích dokumentů náhodnost

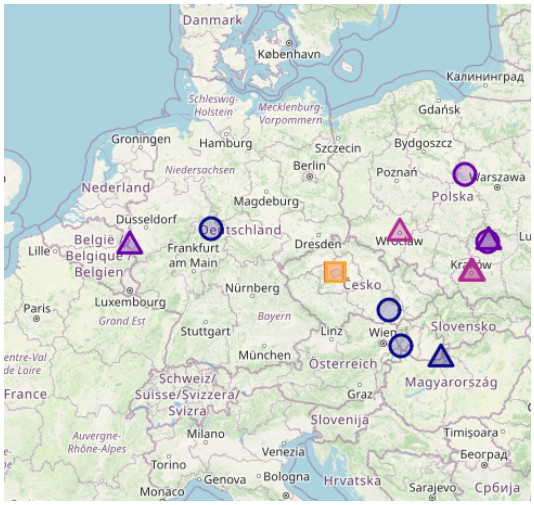
v odpovědích. DBSCAN se ukázal jako nevhodný pro vysokou míru klasifikace vzorků jako šumných.

Popularita používání modularity jako principu pro detekci komunit, jejímž zástupcem je právě Louvein, čelí kritikům mezi jaké patří např. Peixoto (2023). Peixoto ovšem nepředkládá pouze kritiku, ale nabízí srovnání přístupů využívajících modularity s alternativou v podobě Bayesovského modelování – *stochastic block model* (viz Zhang a Peixoto (2020)).

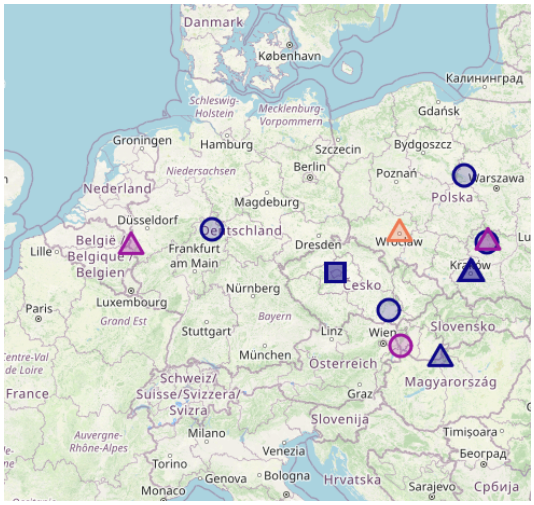
Podobným směrem míří i modelování témat, s pomocí kterého se nám však nepovedlo dosáhnout uspokojivých výsledků. Faktem je, že jsme nakonec, i přes potenciál získané pravděpodobnostní distribuce, dělali „tvrdá“ rozhodnutí (volba maxima). Není ovšem jasné, jak by se s distribucí vlastně dalo či mělo pracovat. Dalším problémem v našem použití může být nízký počet dokumentů v kolekci na rozumné modelování (lingvisté typicky užívají řádově větší sady).

Zajímavý je výstup pokusu představeného v sekci 4.4. Poznání, že pro zkoumání tradic je potřeba snížit granularitu z pramenů na nižší úroveň, nás posouvá k dalším výzkumným krokům ve směru porovnávání výsledků jednotlivých svátků. Svátky nám prameny dělí na uchopitelné jednotky a v případě, že budeme zkoumat, v rámci jaké podmnožiny svátků se určité prameny (či přesněji jejich části odpovídající daným svátkům) chovají obdobně, lze stále hovořit o hledání tradic.

Z nápadů na pokračování zmiňme přístup pro zvýšení jistoty, při kterém by se pro zvolená data našly podmnožiny v rámci komunit, které jsou stabilní napříč metodami. Zkoumali bychom sice pouze část pramenů (pokud by nebyla shoda nalezených rozdělení úplná), ale zato takové, na kterých se více přístupů shodlo, že patří k sobě. Pro repertoár ke sv. Vojtěchu, jehož dvě možná rozdělení ukazují obrázky 4.17 a 4.18 by se jednalo o 7 ze 13 pramenů.



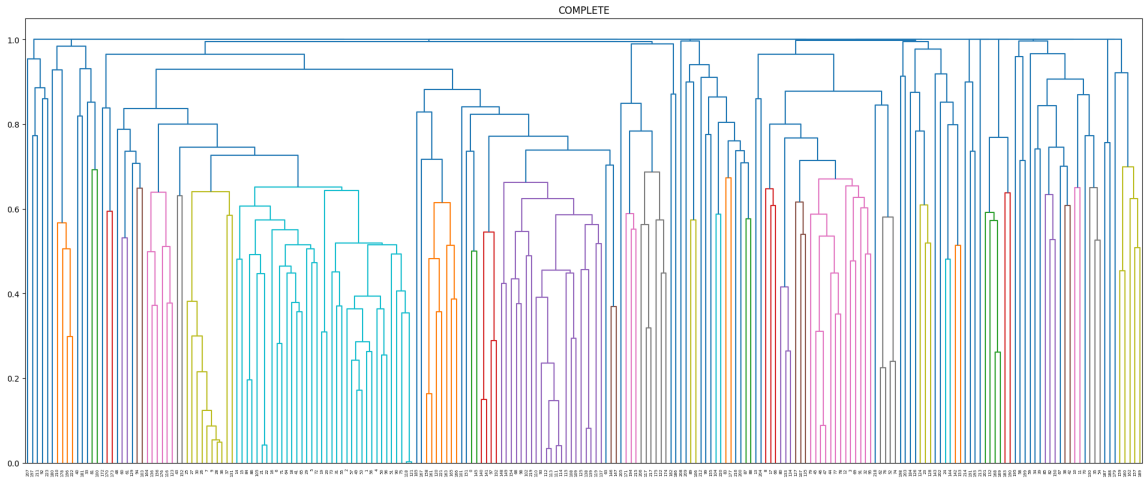
Obrázek 4.17: Komunity pro sv. Vojtěcha nalezené algoritmem Louvein při použití Jaccardovy vzdálenosti



Obrázek 4.18: Komunity pro sv. Vojtěcha nalezené skrze tematický model pro 20 témat

Mezi další možné kroky lze zařadit také přeměření vlivu *cursu* a geografie na výsledky algoritmu Louvein, jakožto nejméně náhodně působícího přístupu. Vzhledem k řídkosti dat a neprůkaznosti výsledku by znovu bylo vhodné přizvat hudební experty.

Z klasických metod shlukové analýzy mohou v našich pokusech vypadat opomenuté hierarchické přístupy (vedle algoritmu k-means, jehož nevýhodou je nutná znalost počtu témat a který standardně pracuje se souřadnicemi bodů). Následující schéma 4.19 ukazuje jeden ze sestavených dendrogramů. Nebylo ovšem vůbec jasné, jaký z přístupů pro stavbu dendrogramu užít (*complete*, *average* či *ward*) a hlavně, ve kterém bodě strom „zařízнout“ a komunity číst.



Obrázek 4.19: Uklázka dendrogramu pro všechna data, přístup *complete*

V rámci evaluace by mohla být možnou zlepšující cestou snaha sehnat od muzikologů více podobných referenčních bodů, jako může být představených šest českých pramenů od prof. Davida Ebena, a provádět evaluaci vůči nim. Ovšem, jak jsme naznačili i výše v části 4.2.2, jedná se o netriviální problém i pro odborníky, a navíc následný způsob práce s takovou informací také není zcela jednoznačný.

V otázce existence repertoárových tradic jsou představené experimenty neuspokojivé. Příčin může být několik. Je možné, že jsme nezvolili vhodné metodické přístupy – ať už obecně, tak v kontextu dat. Jako vhodné potenciální pokračování práce se nám v tomto směru jeví výše zmiňovaný *stochastic block model*. Dále se potýkáme s potenciálním nedostatkem dat (viz 3.3), kde, i přes pocitově obsáhlou databázi kompletovanou mnoha lidmi již mnoho let, máme v porovnání se středověkou realitou stále jen zlomek dat, navíc nerovnoměrné distribuce. Přenosy repertoáru skutečně mohou mít natolik komplikované cesty, že je námi použitá globální (ptačí) perspektiva nezvládá uchopit. Přesto vidíme přínos provedené práce v ukázce (ne)fungování představených tří metod a tedy v prvním probrání a protřízení alespoň části dostupného výpočetního aparátu.

Pohled na celou problematiku a finální odsouzení či pochválení metod by bylo, obzvlášť z důvodu chybějících evaluačních dat, vhodné dělat s dohledem muzikologa znalého mnohých zákoutí materiálu, potenciálních spojnic, historických vývojů a některého kusu materiálu podrobně. Proto je pro takové, kteří jsou ochotní si v prvním kroku svého bádání pomoci technikou, připraven nástroj, poskytující možnou základní (hrubou) analýzu a také geografické vizualizace. (Právě digitální gregorianistika má v rámci muzikologie nezanedbatelnou tradici a například projekt DACT[8] sdružuje více než dvacet institucí, jejichž členové o takovéto nástroje pro digitální gregorianistiku zájem mají.)

[8] `https://dact-chant.ca/`

5. Uživatelská dokumentace

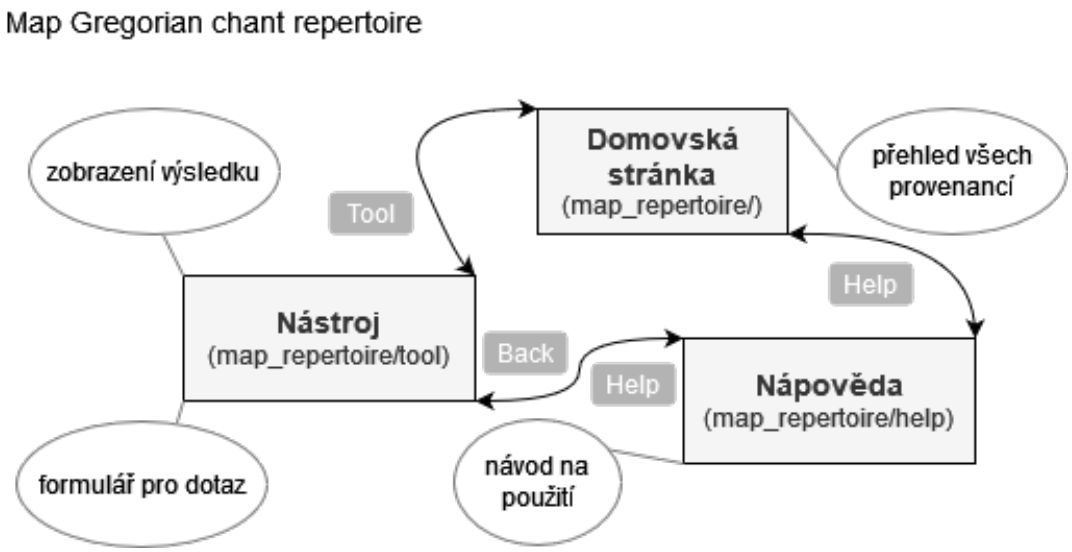
Jako další část práce představujeme nástroj pro muzikology umožňující vizualizovat výsledky algoritmů popsaných v části 4. Jeho cílem je také přilákat hudební odborníky ke strojovému výzkumu a nabídnout jim několik metod pro hledání komunit či tradic, jejichž výsledky posléze mohou podrobit podrobnějšímu zkoumání.

Jedná se o lokální webovou aplikaci a lze k ní po patřičném nastavení přistupovat skrze webový prohlížeč. Po spuštění poběží na adrese `http://127.0.0.1:8000/`.[1] Kroky potřebné k instalaci a spuštění jsou popsané v příloze A.2.

Cílem této kapitoly je popsat funkcionality nástroje a také poskytnout návod k jeho používání.

5.1 Rozvržení stránek

Celý nástroj běží na třech stránkách, jak je naznačeno ve schématu 5.1. Stránky jsou v šedých obdélnících, jejich funkce v bílých oválech, černé šipky značí možné přechody (stránky jsou bezezbytku vzájemně průchozí všemi směry, přechody na domovskou stránku lze provést pomocí kliknutí na název aplikace v šedé liště v horní části okna).



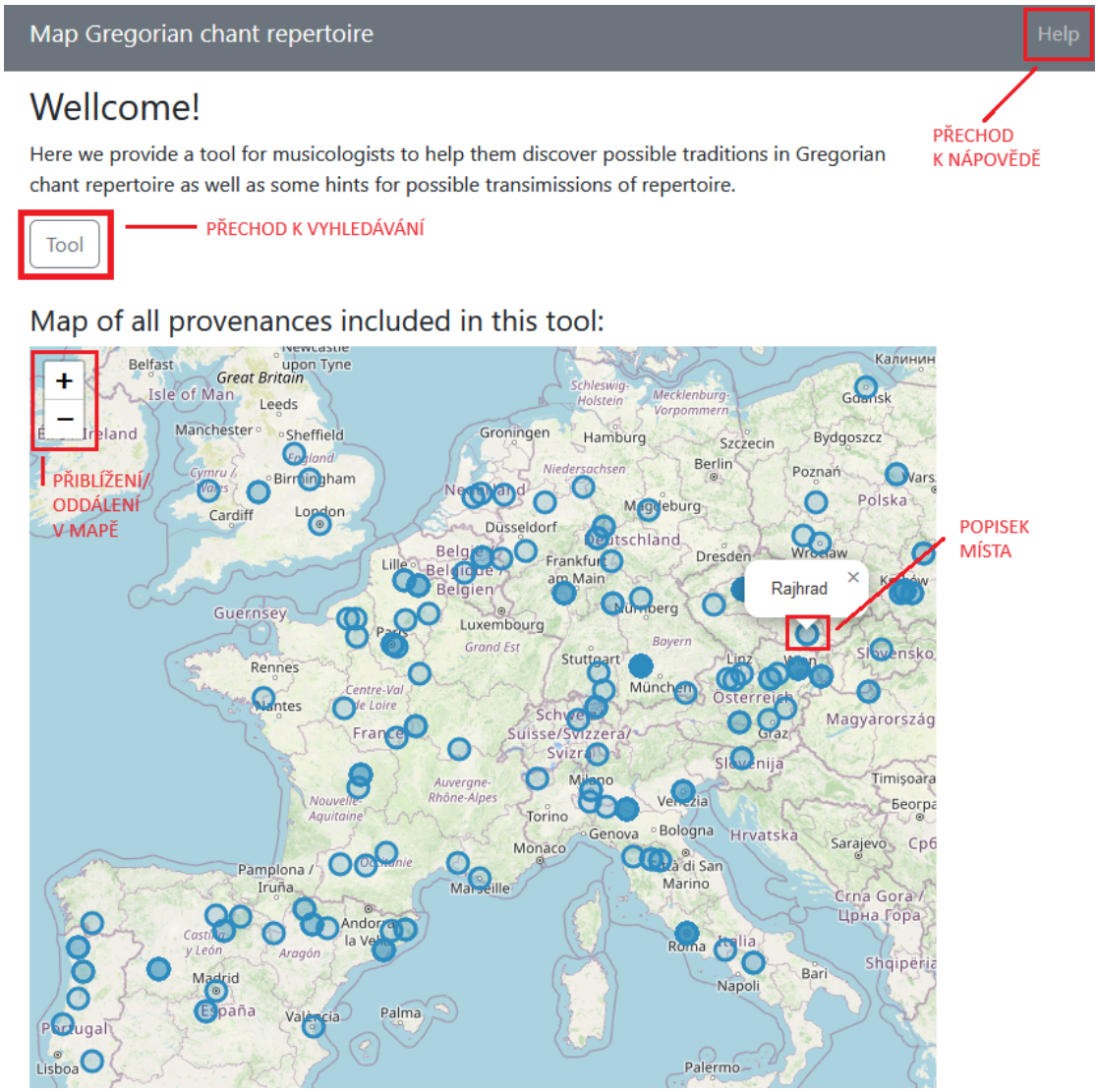
Obrázek 5.1: Schéma částí webové aplikace

5.2 Domovská stránka

Úvodní nebo též domovská stránka poskytuje pár slov o nástroji, přístup ke stránce nápovědy a také mapu všech pramenů, jejichž provenance je nástroji známá.

[1] Defaultní adresu lze změnit zapsáním jiné adresy za spouštěcí příkaz.

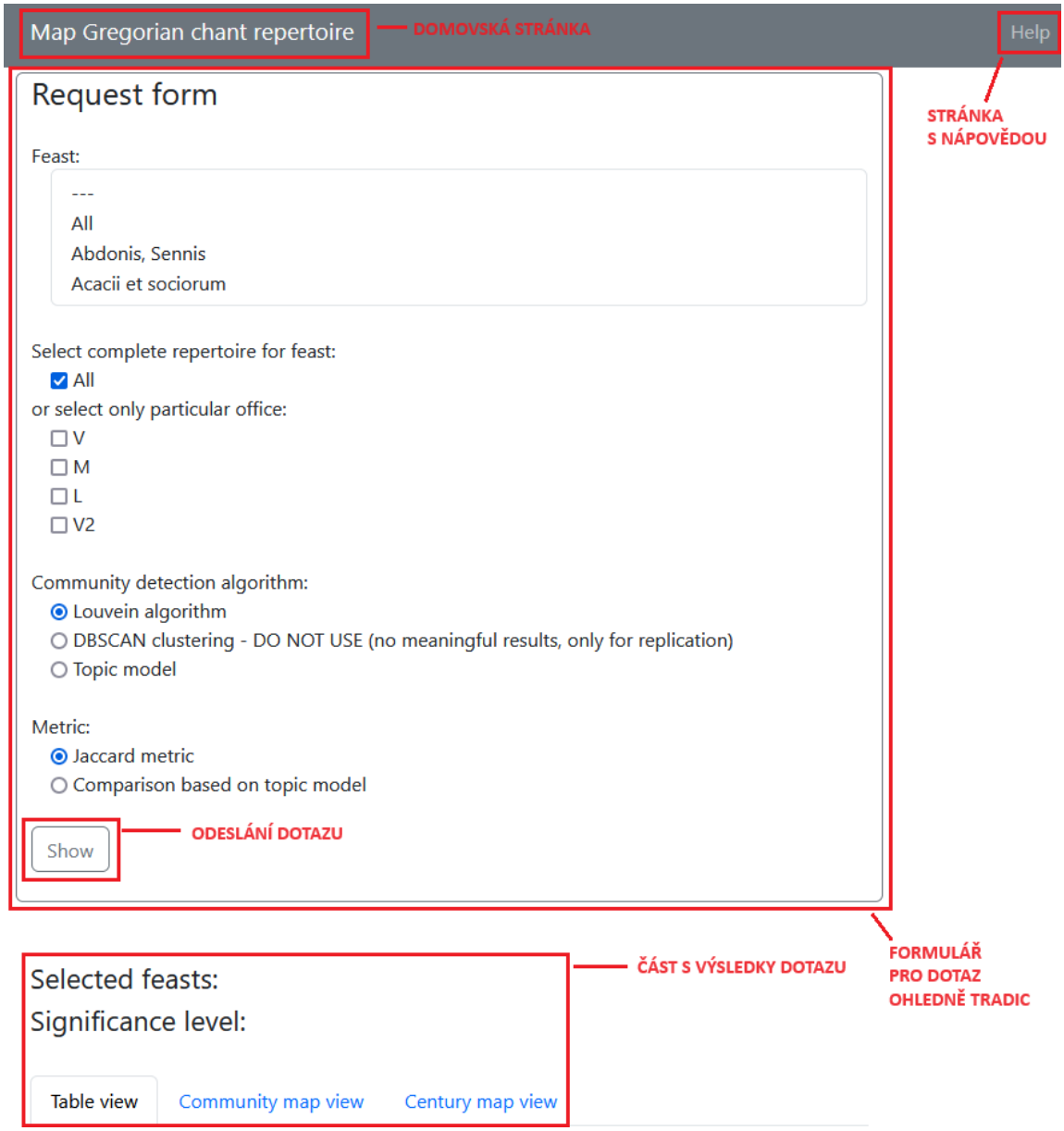
Mapa tu není jen k chlubení, ale také k vytvoření představy o tom, co lze nebo naopak nelze od nástroje v odpovědích očekávat a jak nerovnoměrná je distribuce poskytovaných pramenů. Tedy, že pokud se nám zdá, že chybí očekávaná vazba do některého státu, může to být prostě tím, že data k němu nejsou v rámci nástroje k dispozici atp. Podoba úvodní stránky spolu s popisky efektů kliknutí na jednotlivé části je v náhledu 5.2.



Obrázek 5.2: Náhled domovské stránky aplikace

5.3 Stránka s nástrojem

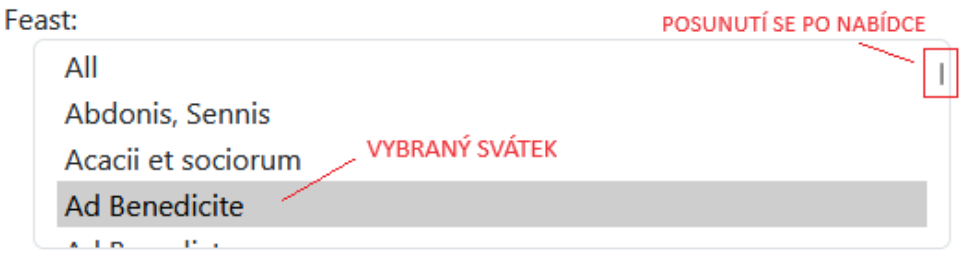
Pro vznesení dotazu je třeba se přesunou na stránku nástroje (tlačítkem Tool, naznačeno v 5.2). Zde dochází k zadávání dotazů a také následně k zobrazování výsledků. Náhled stránky vidíme v 5.3.



Obrázek 5.3: Náhled stránky s nástrojem

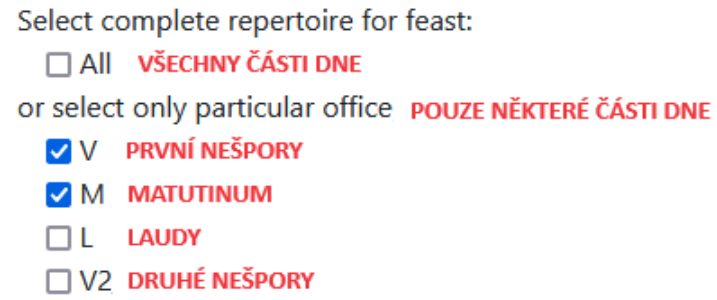
Vyplnění formuláře se sestává ze čtyř kroků:

- **Výběr jednoho a více svátků** z posuvné nabídky (případně možnost All - všechny)



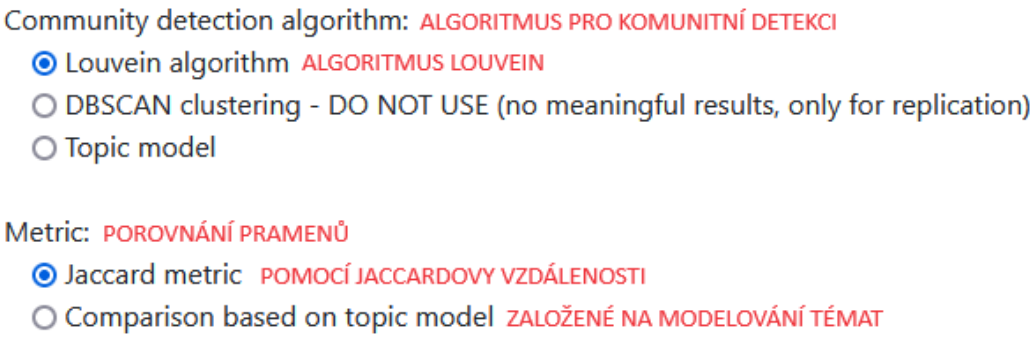
Obrázek 5.4: Výběr svátku či svátků pro výpočet

- **Výběr zahrnutých oficií** (použité kódy viz příloha A.1)



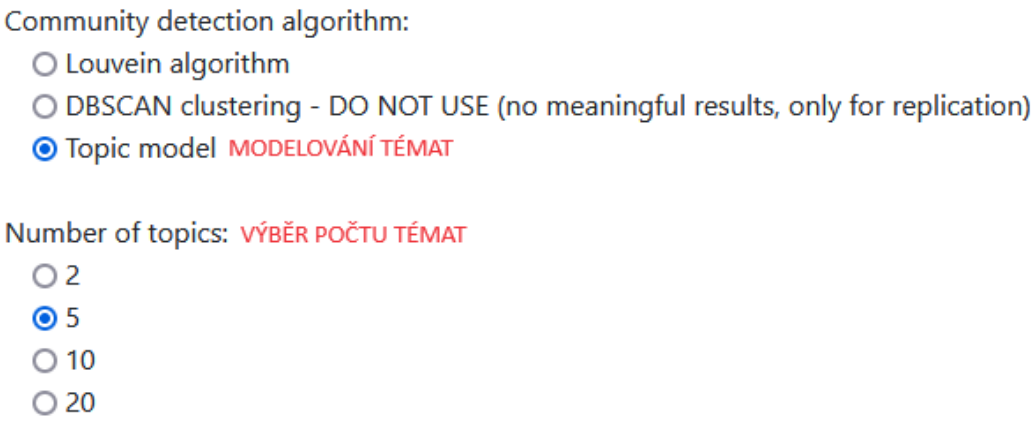
Obrázek 5.5: Výběr oficií zahrnutých do výpočtu

- **Výběr algoritmu** a jeho případné další vlastnosti
  - Algoritmus Louvein (viz 4.1.2) a možnosti porovnávání pramenů (viz 4.1.3)



Obrázek 5.6: Výběr algoritmu Louvein a jeho možnosti

  - Algoritmus DBSCAN - není vhodný k výzkumu
  - Výpočet s pomocí modelování témat (LDA)



Obrázek 5.7: Výběr modelování témat a jeho možnosti

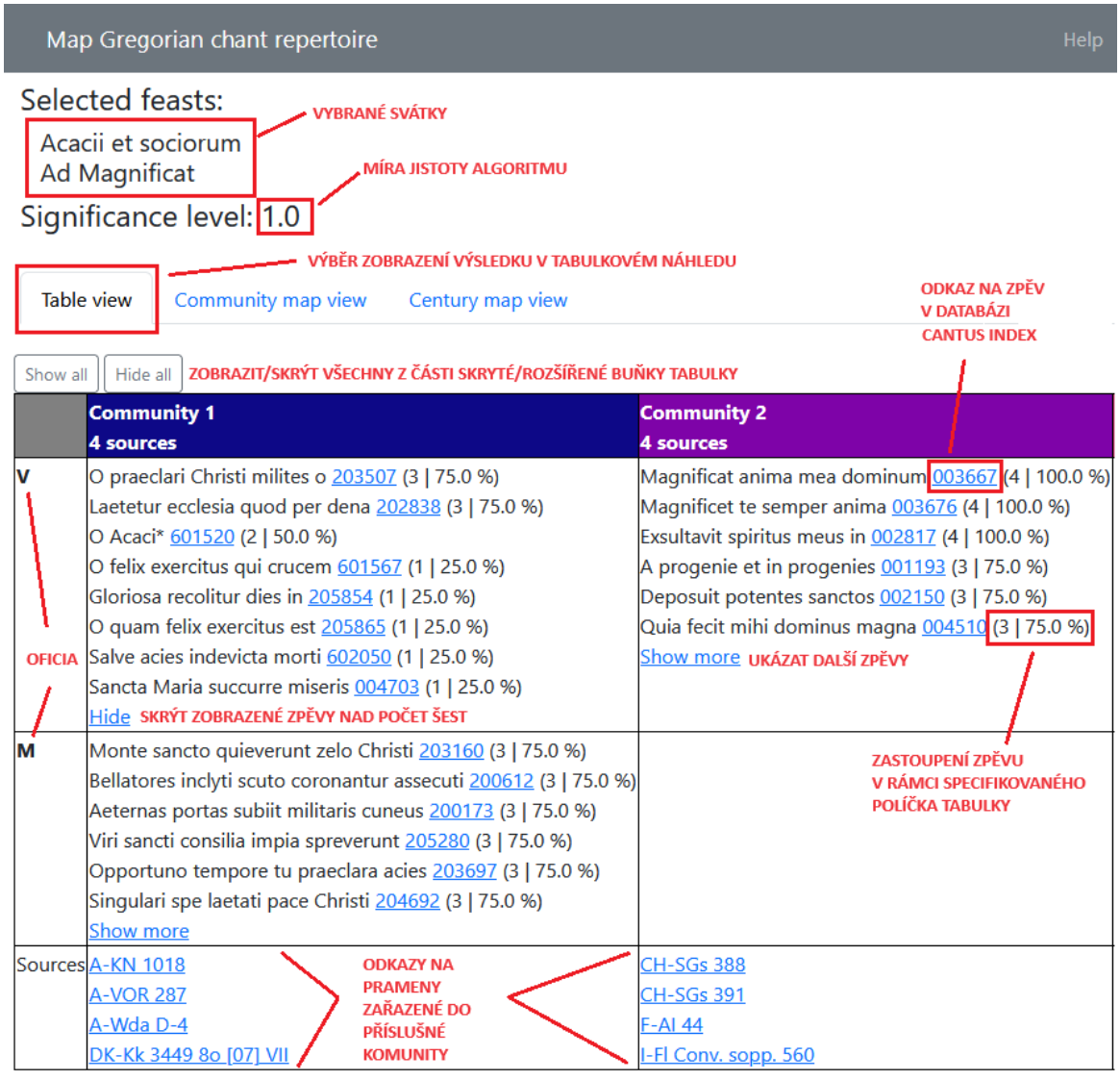
Po vyplnění je ještě potřeba formulář odeslat, a to pomocí tlačítka Show (naznačeno v náhledu 5.3).

5.3.1 Výsledky

Po odeslání dotazu dojde k zobrazení výsledku v dolní části stránky, jak je naznačeno v náhledu 5.3. Délka čekání na odpověď se liší podle počtu svátků a zvoleného algoritmu. Varianta všechny svátky (All) při volbě porovnávání pramenů s pomocí tematického modelování může trvat až půl minuty (záleží také na výkonu počítače).

Nástroj umožňuje zobrazení výsledku ve třech náhledech:

- **Tabulka**
  - Poskytuje informace o pramenech v každé z komunit.
  - Výpočet míry jistoty je popsán v sekci 6.2.1.



Obrázek 5.8: Podoba výsledků v tabulce.

– Její rozvržení vychází z formátu tabulky v existujícím nástroji Cantus Analysis Tool,[2] který funguje při databázi Cantus Index.

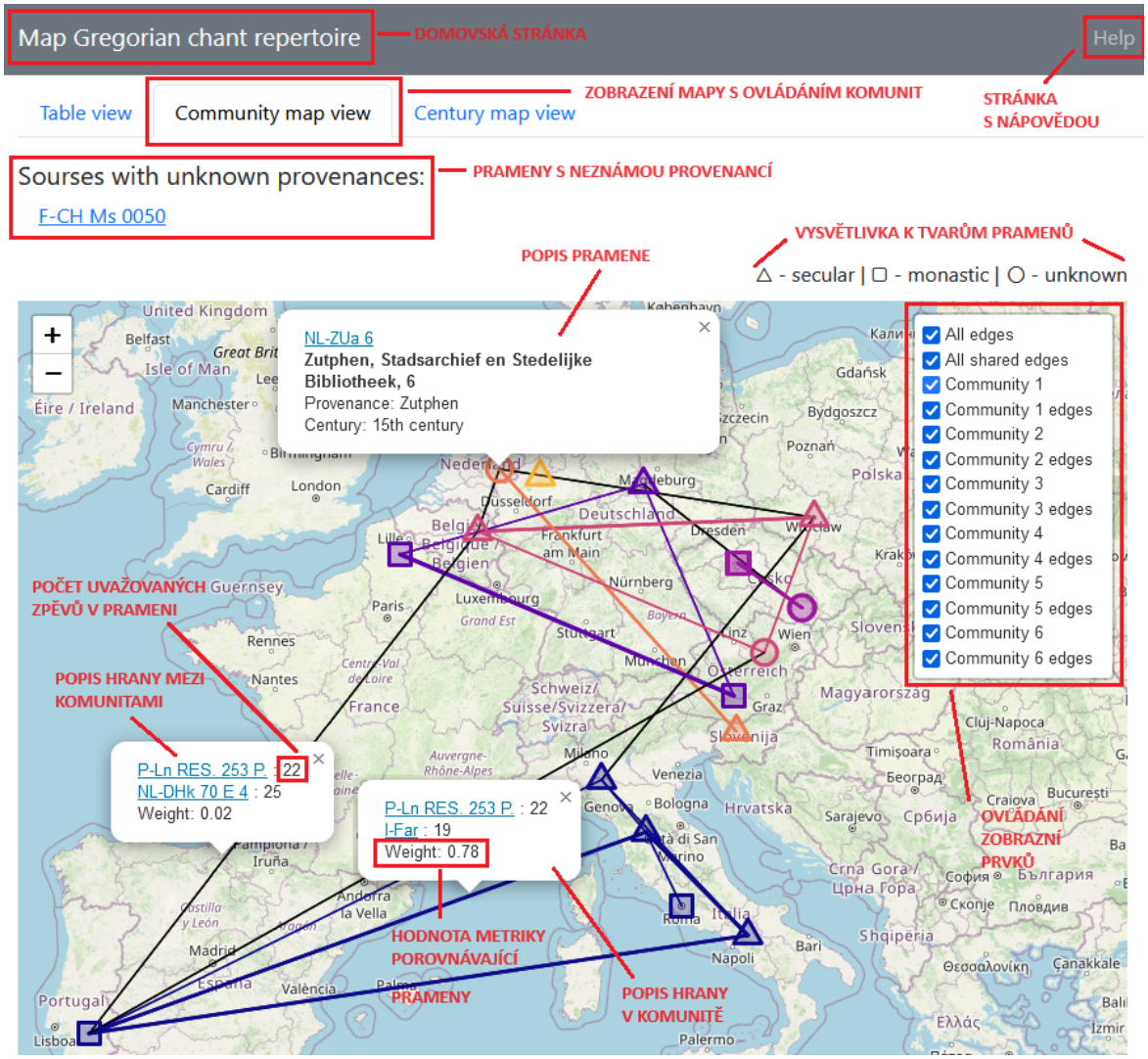
– Každá buňka v těle tabulky zobrazuje zpěvy z knih v dané komunitě (sloupec) odpovídající zvoleným svátkům a příslušnému oficiu[3] (řádek).

– Frekvence u zpěvů je počítána pro knihy v komunitě pro všechny vybrané svátky pro dané oficium tak, že za každou knihu je zpěv započítán pouze jednou.

- **Mapa po komunitách**

  – Prameny jsou v mapě reprezentovány geometrickými obrazci, barva odpovídá zařazení do komunity (stejná barva je i v záhlaví tabulky).

  – Hrany v rámci komunity mají barvu komunity.

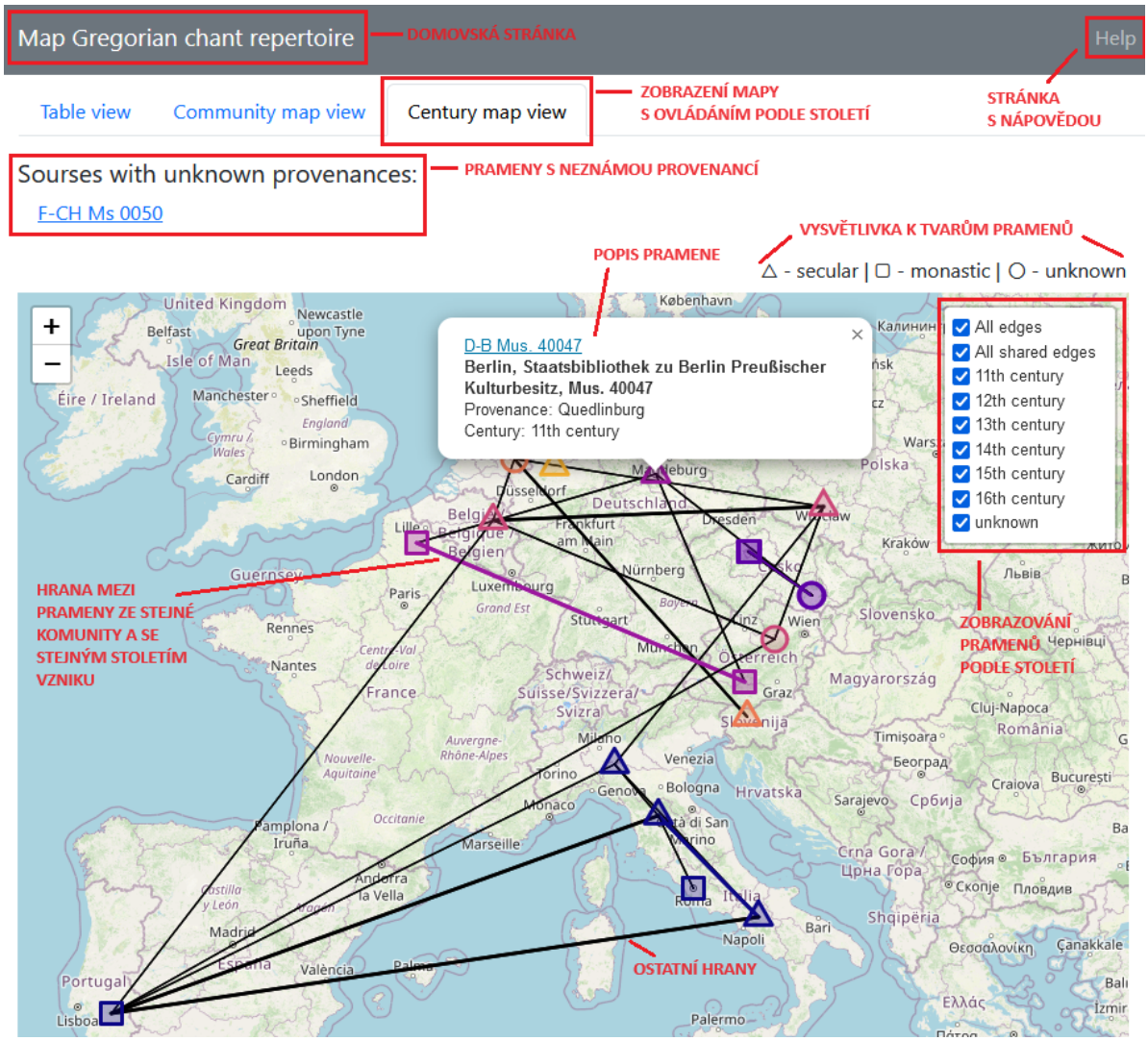


Obrázek 5.9: Podoba mapy po komunitách

[2] `https://cantusindex.org/analyse`
[3] Kódy oficií lze nalézt v příloze A.1.

- Hrany mezi prameny z různých komunit mají černou barvu.
- Na prameny i na hrany lze kliknout a zobrazit si rozšiřující informaci.
- V případě, že mezi vrcholy (prameny) nevede hrana, prameny nesdílejí žádné zpěvy.

- **Mapa po stoletích**
  - Prameny jsou v mapě reprezentovány geometrickými obrazci, barva odpovídá zařazení do komunity (stejná barva je i v záhlaví tabulky).
  - Hrany v rámci komunity, které pocházejí ze stejného století, mají barvu komunity.
  - Ostatní hrany jsou černé.
  - Na prameny i na hrany lze kliknout a zobrazit si rozšiřující informaci.
  - V případě, že mezi vrcholy (prameny) nevede hrana, prameny nesdílejí žádné zpěvy.



Obrázek 5.10: Podoba mapy po stoletích

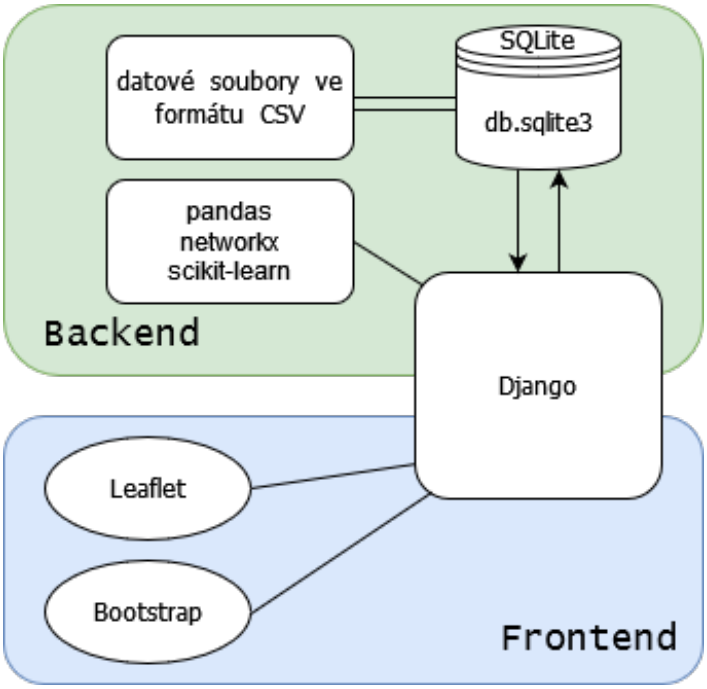
6. Vývojová dokumentace

Projekt je webová aplikace ve frameworku Django.[1] Kroky potřebné k její instalaci a použití jsou popsány v příloze A.2. Software lze rozdělit na dva základní celky – backend a frontend. Jejich konkrétní funkce a interakce v našem nástroji naznačuje schéma 6.1. Podrobné popsání příslušných částí, jejich propojení a fungování je předmětem této kapitoly.

Backend je ta část aplikace, která běží v pozadí, má na starosti práci s daty a databází a počítání, které by bylo pro frontend příliš výpočetně náročné, resp. které mají referenční implementaci v jazyce Python, který se na backend používá. V našem případě je jeho úkolem nalézt repertoárové komunity a vrátit frontendu jejich vhodnou reprezentaci.

Frontend naproti tomu poskytuje uživatelské rozhraní. Skrze něj dochází k dotazům a zároveň k zobrazení backendem vráceného výsledku. V námi představeném nástroji jsou role frontendu zobrazení dotazovacího formuláře a práce na sestavení tabulky a map s výsledky. Komunikace s uživatelem probíhá skrze webový prohlížeč. Při lokálním spuštění webové aplikace leží domovská stránka na defaultní URL adrese Djanga: `http://127.0.0.1:8000/`.

Aplikace je naprogramovaná ve frameworku Django,[2] v jazyce Python. Backend pak využívá řadu jeho knihoven. Frontend je řešený přes šablonový systém Djanga, avšak používá javascriptový framework Bootstrap[3] pro vizuální podobu stránky a knihovnu leaflet[4] pro mapovou funkcionalitu.



Obrázek 6.1: Schéma rozvržení aplikace

[1] V současnosti není vyřešený deployment, takže je určena k lokálnímu spouštění, avšak počítá se se zveřejněním v rámci projektu DACT (`https://www.dact-chant.ca`).

[2] `https://www.djangoproject.com/`

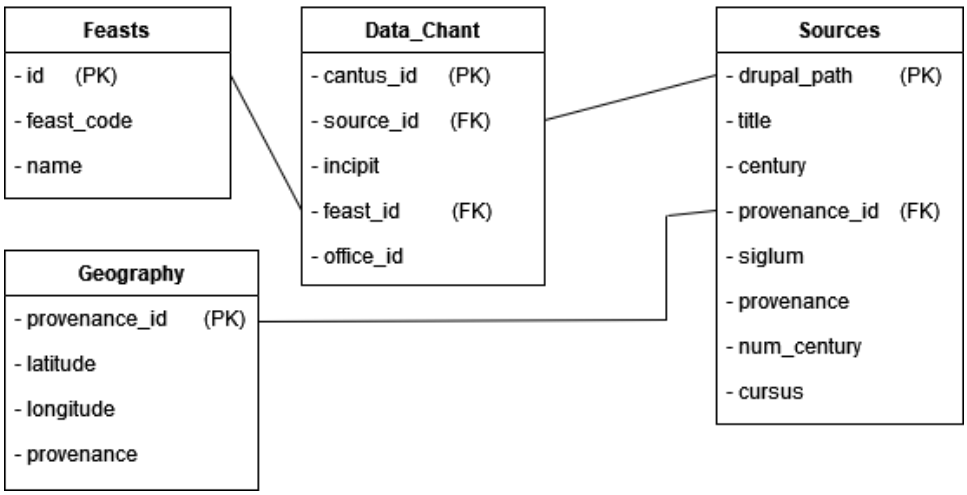
[3] `https://getbootstrap.com/`

[4] `https://leafletjs.com/`

6.1 Data a jejich uložení

Pro ukládání dat používáme databázovou technologii SQLite.[5] Jedná se o odlehčený a samostatný databázový engine, který je jednoduchý na použití. Máme k dispozici předdefinovanou databázi navrženou pro backend. Podstatnou nevýhodou SQLite je nízký výkon. Protože však náš případ použití není tak náročný na tok dat, nepovažovali jsme za nutné defaultní řešení poskytované Djangem měnit.

Tabulky uložené v databázi odráží datové soubory představené v kapitole 3.1, kde ovšem ne všechny jejich sloupečky používáme a tedy do databáze nahráváme. Která data ukládáme a jaká je logika propojení naznačuje schéma 6.2.



Obrázek 6.2: Schéma tabulek v databázi (PK = primární klíč, FK = cizí klíč)

Předpisy jednotlivých tabulek ze schématu 6.2 jsou v souboru `models.py` v rámci naší aplikace. Každá tabulka je třída typu `Model` a jednotlivé sloupce tabulky jsou její atributy. Kromě jednotlivých polí a jejich datových typů popisují třídy také chování v případě chybějící hodnoty (přijímá-li dané pole hodnotu `null`).

Přesun dat z příslušných CVS souborů (viz 3.1) do databáze probíhá spuštěním skriptu `load_csv.py`, který s pomocí knihovny `pandas`[6] dané soubory načte a uloží skrze modely do databáze (`db.sqlite3`). Jestliže došlo k aktualizaci v souboru `sources-of-all-ci-antiphons_OPTIONAL-CENTURY.csv`, pak je potřeba před `load_csv.py` spustit ještě skript `new_csv.py`, který změny aktualizuje v souboru `sources-with-provenance-ids-and-two-centuries.csv` (namapování názvu provenance s provenance_id či případné zavedení unknown, vyplnění kolonky num_century, filtrace pramenů podle velikosti).

[5] `https://www.sqlite.org/`

[6] `https://pandas.pydata.org/`

6.2 Backend

Jak již bylo zmíněno výše, pod pojmem backend rozumíme tu část aplikace, která je zodpovědná za datovou správu (databáze) a výpočty. Tuto část software jsme se rozhodli vytvořit s pomocí jazyka Python. Zvolili jsme si jej, neboť má dostupné a snadno použitelné referenční implementace všech metod, které chceme používat, a také protože je to programovací jazyk v komunitě *digital humanities* nejrozšířenější.[7]

Základem backendu je webový framework Django v jazyce Python. Umožňuje psát webové aplikace rychle a s důrazem na aplikaci samotnou, zatímco se stará o jednotvárné (únavné) části webového vývoje, pro které nám poskytuje přístupnou abstrakci. Každý Djnago projekt je složen z částí zvaných `app` (aplikace). Jedná se o jednotlivé samostatné programové kusy, kde každý poskytuje svoje funkce. Lze je díky modularitě Djanga snadno přesouvat mezi různými projekty. V našem případě máme v projektu Gregorian_chant_repertoire pouze jednu `app`: `Map_repertoire`. Její zapojení do projektu je uvedeno v souborech `settings.py` a `urls.py` projektu.

Podobně jako tabulky databáze, kde se jedná o objekty třídy typu `Model`, má Django řešené i formuláře. Objekt typu `Form` reprezentuje formulář a specifikuje jednotlivá pole, jejich typy, hodnoty u polí s výběreme, defaultní hodnoty a případnou (ne)povinnost vyplnění. V naší aplikaci je pouze jeden formulář, `InputForm`, jehož popis je v souboru `forms.py`. Hodnoty z něj se čtou v rámci funkce `tool` ve `views.py`.

6.2.1 Výpočty v backendu

Výpočetní část backendu se sestává z funkcí ve třech souborech:

- `communities.py`
  - hlavní funkce `get_communities`
  - zpracovává volby uživatele získané z formuláře – které svátky, kterým algoritmem a která oficia
  - vyžádá si příslušná data z databáze a pomocí zvoleného algoritmu najde komunity
  - vrací zkonstruované komunity, váhy hran mezi vrcholy a míru stability (*significance_level*)

- `table_construct.py`
  - hlavní funkce `get_table_data`
  - pro nalezené komunity s pomocí databáze konstruuje obsah tabulky výsledku (viz 5.3.1) v datové struktuře (slovník), jakou lze dobře zpracovat v rámci frontend skrze Django template language

[7] Konkurovat by mu možná mohl jazyk R, ale my potřebujeme více než statistické výpočty.

- `map_data_construct.py`
  - hlavní funkce `get_map_data`
  - na základě nalezených komunit s pomocí dotazování se databáze chystá datovou strukturu (slovník) pro funkce v jazyce Javascript vytvářející mapy

Získávání komunit naprogramované v `communities.py` využívá stejné implementace jako výzkumná část práce (viz kap. 4). Algoritmus Louvein je z knihovny `networkx`, přičemž parametr rozlišení je roven 1.0. V případě této volby uživatele je sestaveno deset verzí komunit, spočítán `sig_level` (Jaccard index (viz 4.2.1), kde při výpočtu je k čitateli i jmenovateli přičten počet pramenů v dotazu), a je vrácena první z nalezených komunit. To může vést k nedeterministickému chování, kdy při opakovaném zadání stejného dotazu jsou odpovědi aplikace rozdílné. DBSCAN je z knihovny `scikit-learn`, ovšem jeho použití pro zkoumání komunit se ukázalo jako nevhodné. Pro modelování témat jsou v nástroji uložené čtyři dvojce modelů pro 2, 5, 10 a 20 témat. Jedná se vždy o *CountVectorizer* vybraný vzhledem k počtu témat (viz 4.3.3) a *LDA* trénované na celé datové sadě (250 pramenů, všechny svátky). V případě, že uživatel žádá komunity nalezené s pomocí tematického modelování, pak:

1. se sestaví dokumentová struktura pro data žádaných svátků,
2. načtou se uložené modely žádaného počtu témat,
3. dokumenty se s pomocí modelu *CountVectorizer* přetransformují do matice frekvencí slov v dokumentech
4. načtené *LDA* následně vymodeluje tematické distribuce a
5. podle nich (bereme téma s nejvyšší pravděpodobností) se vrátí komunity.

Pro variantu hledání komunit skrze modelování témat není `sig_level` z čeho počítat, kód vrací '–'.

6.3 Komunikace hlavních komponent

Průběh volání funkcí popsaných v sekci 6.2.1 i následná komunikace mezi backendem a frontendem je zapsaná v souboru `views.py`. V něm nalézáme tři funkce, z nichž každá slouží pro jednu ze tří stránek v aplikaci (`index`, `tool`, `help`). Funkce bere žádost webu (*web request*) a vrací *web response* – v našem případě obsah stránky k zobrazení, např. `return render(request, "map_repertoire/tool.html", context)` pro zobrazení stránky `tool`, kde `render` vyvolá načtení nové stránky uživateli ze zvoleného HTML souboru. Proměnná `context` v návratovém volání funkce je slovník, který slouží k předání dat mezi backendem a frontendem.

Informace předávané z backendu do frontendu pro zobrazení `tool.html`:

- `feasts` - seznam jmen žádaných svátků
- `form` - odpovědi uživatele ve formuláři

- `sig_level` - míra, jakou si je algoritmus jistý výsledkem
- `map_data` - datová struktura popisující zobrazení výsledku v mapě
- `tab_data` - datová struktura popisující obsah výsledkové tabulky

Tyto pak frontend používá v rámci souborů v HTML a v jazyce Javascript pro zobrazení stránky `tool` po odeslání formuláře uživatele. Stránka `help` nevyžaduje od backendu nic. Stránka `index` (úvodní, domovská stránka aplikace) potřebuje proměnnou `map_data_all`, aby zobrazila všechny nástroji známé body na úvodní mapu.

6.4 Frontend

Frontend je část programu zodpovídající za uživatelskou část aplikace – sběr žádosti uživatele a následné zobrazení výsledků hledání komunit. Data k zobrazení jí poskytuje ve vhodné formě backend. Komunikace těchto dvou částí probíhá skrze funkce ve `views.py`.

Průběh jednoho požadavku od uživatele (odeslání validního formuláře) vypadá následovně:

- Data z formuláře, skrze nějž uživatel specifikuje svůj požadavek, se sesbírají ve `views.py`.
- Backend na jejich základě vytvoří `sig_level`, `map_data` a `tab_data` a vloží je do slovníku `context`.
- Dojde k vyvolání nového načtení `tool.html` (`return render(...)`).
- S pomocí Django template languge dojde k iteraci skrze `tab_data` a předchystaná tabulka se tak naplní příslušným obsahem.
- Slovník `map_data` se z proměnné `context` zpřístupní javascriptové části soboru.
- V části `tool.html` psané v jazyce Javascript se zavolá `getMaps(map_data)` ze souboru `static/create_map.js`.
- Získané mapy se spojí s odpovídajícími `<div>` elementy na `tool` stránce.

Vedle stránky s formulářem a se zobrazenými výsledky zajišťuje frontend ještě úvodní stranu se základními informacemi a mapou všech nástroji známých provenancí (`index`), která má skrze `context` od backendu `map_data_all` a volá `getMapOfAllSources()` opět z `create_map.js`. Dále frontend zajišťuje stránku s nápovědou (`help`). Základní kostra všech tří stránek leží v rámci `app` ve složce `templates/Map_repertoire`. Všechny rozšiřují základní `base.html` soubor ze složky `templates`, který obsahuje společné části stránky a popisy chování CSS elementů.

K vytvoření map ve funkcích `getMapOfAllSources` a `getMaps` používá aplikace knihovnu `leaflet` jazyka Javascript (verze 1.9.4). Jedná se o základní knihovnu pro tvorbu interaktivních map. (Existuje knihovna `folium` jazyka Python,

která umožňuje používat `leaflet` z prostředí Pythonu, ovšem brzy jsme při vývoji s ní narazili na chybějící možnosti způsobené nekompletním pokrytím původní knihovny.) Soubory knihovny `leaflet` jsou uložené v aplikaci lokálně ve složce `static/leaflet`.

Pro jiné něž kruhové tvary bodů v mapě (trojúhelník a čtverec) používáme rozšíření `Leaflet.SvgShapeMarkers`.[8] To poskytuje tvary pro zobrazení bodů na SVG vrstvy. Ovšem v pokročilejší fázi vývoje jsme zjistili, že v případě velkého množství vrcholů (a tedy i hran) na mapě, přičemž vrcholů může být až okolo dvou set a hran tedy až čtyřicet tisíc, dochází k výraznému zpomalení interaktivity mapy.[9] Rozhodli jsme se proto přesunout k zobrazování na technologii Canvas. Knihovna `leaflet` tuto variantu podporuje (hrany a kruhové body fungovaly), ovšem k rozšíření používanému pro další tvary vrcholů bylo třeba dopsat vhodnou podporu. Rozšiřující kód je zapsán za původní částí kódu pro SVG v souboru `static/leaflet-svg-shape-markers.js`.

Pro pozvednutí vizuální podoby aplikace nad základní HTML design jsem zvolili CSS framework `bootstrap` (ve verzi 5.3). Jedná se o snadnou a přístupnou variantu poskytující vylepšení designu stránky, které by nemělo komplikovat práci ani případnému budoucímu jinému programátorovi. Kód k frameworku (CSS a Javascript soubory) je opět uložen lokálně, a to ve složce `static/bootstrap`.

6.5 Závislosti

Běh aplikace a její funkce závisí na několika dalších programech. Pro běh je nutné mít nainstalovaný Python. Při vývoji byla použita verze 3.11.6, není tedy garantováno, že aplikace poběží s nižšími verzemi. Dále je potřebná instalace Django (verze 5.0), `pandas` (verze 2.1.2), `networkx` (verze 3.3), `matplotlib` (verze 3.8.2), `scipy` (verze 1.12.0), `numpy` (verze 1.26.1) a `scikit-learn` (verze 1.4.2). Pro instalaci všeho potřebného postupujte podle instrukcí v příloze A.2.

[8] `https://github.com/rowanwins/Leaflet.SvgShapeMarkers.git`

[9] Možnost zobrazování všech hran, nikoliv pouze těch uvnitř komunit, má význam např. při zkoumání jednoznačnosti komunit.

Závěr

Gregoriánský chorál je jako středověký kulturní a hudební fenomén předmětem zájmu a studia mnoha muzikologů. Náš interdisciplinární výlet, podpořený jejich digitalizačními snahami, přináší jiný druh výsledků, než na jaký je tento obor zvyklý. Namísto budování přehledu na základě mnoha detailů vytváříme onen přehled pomocí výpočetních modelů, respektive zjišťujeme, jaké předpoklady jsou pro vytvoření takových modelů vhodné. Zároveň je práce netradiční i z hlediska výpočetních pokusů v informatice: pohybujeme se v oblasti, která neposkytuje správné odpovědi, proti nimž by bylo možné jednoznačně změřit přesnost zvolených metod, v oblasti, kde i zpřesnění výzkumných otázek, respektive protřízení předpokladů pro další práci, může být samo o sobě podstatným přínosem.

Výpočetní pokusy představené v kapitole 4 nepotvrdily předpokládanou existenci repertoárových tradic. Dvě ze tří zkoumaných metod se ukázaly jako nevhodné při výpočtech na plné datové sadě (viz kap. 3) i některých jejích částech. Výsledky algoritmu Louvein se jeví nadějně z hlediska stability, jejich obsahovou relevanci je však třeba konzultovat s odborníky. Je také možné, že šíření a přenos repertoáru gregoriánského chorálu (tedy skutečný svět) jsou nejenom nedostatečně popsané, ale také příliš mnohovrstevnaté na to, aby bylo možné je formalizovat společnou parametrizací napříč celým repertoárem.

Kladná očekávání od výzkumu nad datovou sadou z Cantus Index byla dána, s vědomím jejích limitací (viz 3.3), jejím rozsahem, který je u humanitního materiálu nebývalý. Přehledová čísla samotná (viz 3.2), jakkoliv vypadala slibně, byla výsledkem katalogizačních procesů strukturovaných ne zcela vhodně pro podobné výzkumné otázky. Hiley (1993) zmiňuje, že René-Jean Hesbert si pouze na porovnávání responsorií adventních nedělí zvládl shromáždit 800 knih (antfonářů i breviářů) a podobně rozsáhlá porovnání kousků repertoáru prováděl i Le Roux (1961). Vedle toho je 250 pramenů s alespoň 100 zpěvy, což je množství v naší sadě, číslem stále nízkým. Kromě množství dat lze za problém považovat i potenciální nereprezentativnost výběru (viz 3.3). S tím souvisí i to, že ne všechna data jsou se všemi užitečná (zajímá-li mě konkrétní svátek, spoustu zdrojů z porovnání vyřadím), tedy potřeba dalších knih rychle narůstá.

V otázce tradic nad větší částí repertoáru (oproti středověké realitě ovšem stále malém, viz 3.3) i jeho částech (např. deset největších svátků) je třeba přiznat, že jsme s našimi metodami a znalostmi neuspěli. Ovšem povedlo se ukázat, že algoritmus Louvein umí hledat stabilní uspořádání i na repertoáru jednotlivých svátků, přičemž chceme-li konzultovat výsledky s odborníky – lidmi, toto je jednoznačně uchopitelnější přístup. To bylo podpůrnou motivací pro vznik interaktivního nástroje, který by poskytoval přiblížení výsledků výpočetního výzkumu muzikologům. Ten je představený v kapitolách 5 a 6.

V rámci dalších kroků lze uvažovat o provádění měření na mešním repertoáru (naše pokusy používaly výhradně repertoár oficií) a potvrzování či vyvracení předpokladu, že je stabilnější než repertoár oficia. Za prozkoumání stojí také *stochastic block model* zmiňovaný v 4.5, který by mohl, jakožto přístup potenciálně méně náročný na počet pramenů než LDA, být dobrým doplněním k modelování témat. Na pokus popsaný v sekci 4.4 by bylo možné navázat hledáním množin svátků, jejichž tradice jsou navzájem kompatibilní. V situaci, kdy určité prameny pro

určité svátky fungují takto „společně“, tedy korelovaně, se může nějaká regionální/institucionální/politická identita odrážet.

Výzkum ukazuje na zásadní vliv výběru dat na výsledek a jeho platnost. Filtrace vstupních pramenů a zpěvů je tedy krokem složitým a ošemetným. Proto je v rámci vývoje nástroje dalším logickým krokem jeho rozšíření o nahrávání vlastní datové sady pro výpočty, neboť takové řešení poskytuje uživatelům maximální flexibilitu ve výběru dat.

Výpočetní muzikologie v otázce gregoriánského chorálu je zatím nepříliš prozkoumanou oblastí, jak o tom píšeme v kapitole 2. Proto, přestože jsme pocitově nedosáhli žádného „krásného“ výzkumného výsledku, vnímáme i provedení prvotních pokusů, problematizaci některých předpokladů a nástroj pro upřesňování otázek jako přínosy, které snad umožní v mezioborové spolupráci poodhalit některá z tajemství, jež gregoriánská tradice ještě skrývá.[10]

Obrázek 9: Vyobrazení Getsemanské zahrady z breviáře MS M.893, f. 17r

[10]Se středověkým citem pro obrazný jazyk by se dalo říci, že výpočetní muzikologie v oblasti gregoriánského chorálu je polem nezoraným, na jehož ploše lze vedle potenciální úrody nalézt též plevel a že jakkoliv naše zahradnické snahy zatím žádnou lahodnou křupavou mrkev nepřinesly, tak jsme snad alespoň část toho býlí vytrhali a navíc připravili muzikologickým zahradníkům výpočetní motyku na zbytek plevele, který nám se rozpoznat nepovedlo.

Seznam použité literatury

Apel, W. (1958). *Gregorian chant*, volume 601. London: Burns & Oates.

Blei, D. M., Ng, A. Y. a Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Reasearch*, **3**, 993–1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. a Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, (10), P10008. doi: 10.1088/1742-5468/2008/10/P10008. URL `https://dx.doi.org/10.1088/1742-5468/2008/10/P10008`.

Cornelissen, B., Zuidema, W. a Burgoyne, J. A. (2020a). Studying Large Plainchant Corpora Using chant21. In *7th International Conference on Digital Libraries for Musicology (DLfM)*. doi: 10.1145/3424911.3425514.

Cornelissen, B., Zuidema, W. a Burgoyne, J. A. (2020b). Mode Classification and Natural Units in Plainchant. In *21st ISMIR Conference*, page 869–875.

De Coul, T. O. (2021). Dealing with change: the Carthusians and Corpus Christi. *Plainsong and Medieval Music*, **30**(1), 29–53. doi: 10.1017/S0961137121000024.

Eipert, T. a Moss, F. C. (2023a). Poster: Communities in Medieval Troper Networks are Shaped by Carolingian Politics. In *10th International Conference on Digital Libraries for Musicology (DLfM)*. URL `https://dlfm.web.ox.ac.uk/sites/default/files/dlfm/documents/media/poster2023_eipert_communities.pdf`.

Eipert, T. a Moss, F. C. (2023b). Monodikit: A data model and toolkit for medieval monophonic chant. In *10th International Conference on Digital Libraries for Musicology (DLfM)*. doi: 10.1145/3625135.3625145.

Eliade, M. (2006). *Posvátné a profánní*. OIKOYMENH, Praha. ISBN 8072981757.

Ester, M., Kriegel, H.-P., Sander, J. a Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*.

Fernández Riva, G. (2019). Network Analysis of Medieval Manuscript Transmission. Basic Principles and Methods. *Journal of Historical Network Research*, **3**, 30–49.

Hagberg, A. A., Schult, D. A. a Swart, P. J. (2008). Exploring Network Structure, Dynamics, and Function using NetworkX. In *7th Python in Science Conference (SciPy2008)*, page 11–15.

Hajič jr., J., Ballen, G. A., Mühlová, K. H. a Vlhová-Wörner, H. (2023). Towards Building a Phylogeny of Gregorian Chant Melodies. In *24th International Society for Music Information Retrieval Conference*. ISMIR. doi: 10.5281/zenodo.10340442. URL https://doi.org/10.5281/zenodo.10340442.

Hesbert, R. (1963–79). *Corpus Antiphonalium officii: Editum a Renato-Joanne Hesbert*. Corpus antiphonalium officii. Herder, Roma.

Hiley, D. (1993). *Western Plainchant: A Handbook*. Oxford University Press. ISBN 0-19-816289-8.

Hiley, D. (2009). *Gregorian Chant*. Cambridge Introductions to Music. Cambridge University Press, New York. ISBN 9780521690355.

Hoffman, M. D., Blei, D. M. a Bach, F. (2010). Online Learning for Latent Dirichlet Allocation. In *23th Advances in Neural Information Processing Systems*. ISBN 9781617823800. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/71f6278d140af599e06ad9bf1ba03cb0-Paper.pdf.

Hoppin, R. H. (2007). *Hudba stredoveku*. edicia preklady. Hudobné centrum, Bratislava. ISBN 9788088884873.

Hornby, E. (2004). The Transmission of Western Chant in the 8th and 9th centuries: Evaluating Kenneth Levy's Reading of the Evidence. *The Journal of Musicology*, **21**(3), 418–457. doi: 10.1525/jm.2004.21.3.418.

Kestemont, M., Karsdorp, F., de Bruijn, E., Driscoll, M., Kapitan, K. A., Macháin, P. , Sawyer, D., Sleiderink, R. a Chao, A. (2022). Forgotten books: The application of unseen species models to the survival of culture. *Science*, **375**, 765—-769. doi: 10.1126/science.abl7655. URL http://science.org/doi/10.1126/science.abl7655.

Knyazeva, E. (2021). *Komunity a jejich detekce v sociálních sítích*. Bakalářská práce. Univerzita Karlova, Matematicko-fyzikální fakulta.

Lacoste, D. (2012). The Cantus Database: Mining for Medieval Chant Traditions. *Digital Medievalist*. doi: 10.16995/dm.42.

Lacoste, D. (2022). The cantus database and cantus index network. In *The Oxford Handbook of Music and Corpus Studies*. Oxford University Press. ISBN 9780190945442. doi: 10.1093/oxfordhb/9780190945442.013.18.

Lacoste, D., Bain, J. a Koláček, J. Catalogue of Chant Texts and Melodies – inventories of chant sources. URL https://cantusindex.org/.

Lanz, V. (2023). *Unsupervised segmentation of Gregorian chant melodies for exploring chant modality*. Diplomová práce. Univerzita Karlova, Matematicko-fyzikální fakulta.

Le Bomin, S., Lecointre, G. a Heyer, E. (2016). The Evolution of Musical Diversity: The Key Role of Vertical Transmission. *PLoS ONE*, **11**(3), e0151570. ISSN 1932-6203. doi: 10.1371/journal.pone.0151570. URL http://dx.doi.org/10.1371/journal.pone.0151570.

Le Roux, R. (1961). Aux origines de l'office festif: les antiennes et les psaumes de matines et de laudes pour noël et le ier janvier. *Etudes grégoriennes*, pages 65–170.

Ottosen, K. (2008). *The Responsories and Versicles of the Latin Office of the Dead*. BoD – Books on Demand, GmbH. ISBN 9788776911867.

Pacovský, K. (2023). *Život v pražském kláštеře sv. Jiří ve středověku*. Disertační práce. Univerzita Karlova, Filozofická fakulta.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. a Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Peixoto, T. P. (2023). *Descriptive vs. Inferential Community Detection in Networks: Pitfalls, Myths and Half-Truths*, pages 22–29. Elements in the Structure and Dynamics of Complex Networks. Cambridge University Press.

Planchart, A. E., editor (2009). *Embellishing the Liturgy* . London. doi: 10.4324/9781315256665.

Roy, W. G. (2022). Repertoire Communities in American Popular Music, 1900–1949. *Sociological Perspectives*, **65**(5), 929–959. doi: 10.1177/07311214221080992.

Savage, P. E. (2019). Cultural evolution of music. *Palgrave Communications*, **5**(1), 1–12.

Steiner, R. (1985). Local and regional traditions of the invitatory chant. *Studia Musicologica Academiae Scientiarum Hungaricae 27, no. 1/4*, pages 131—-138. doi: https://doi.org/10.2307/902146.

Szabová, K. (2021). *Analytical tools for Gregorian chant*. Bakalářská práce. Univerzita Karlova, Matematicko-fyzikální fakulta.

Treitler, L. (1981). Oral, written, and Literate Process in the Transmission of Medieval Music. **56**(3), 471–491. doi: 10.2307/2847738.

Vozár, Z. (2018). Metadata for the Middle Ages: A Network Analysis of Manuscriptorium.com. In *Historical Network Research Conference*.

Wagner, P. (1911). *Einführung in die gregorianischen Melodien; ein Handbuch der Choralwissenschaft*, volume 1. Druck und Verlag von Breikopf & Härtel, Leipzig.

Zhang, L. a Peixoto, T. P. (2020). Statistical inference of assortative community structures. *Phys. Rev. Res.*, **2**, 043271. doi: 10.1103/PhysRevResearch.2.043271. URL `https://link.aps.org/doi/10.1103/PhysRevResearch.2.043271`.

Seznam obrázků

Seznam tabulek

A. Přílohy

A.1 Seznam liturgií a jejich zkratek

identifikátor	Kód	Popis
office_?	?	nejisté
office_c	C	kompletář
office_ca	CA	capitulum
office_d	D	denní hodinky
office_e	E	antifony k Magnificat a Benedictus (*in evangelio*)
office_h	H	antifony založené na textech *Historia*
office_l	L	laudes
office_m	M	matutinum
office_mi	MASS	mše
office_n	N	nona
office_p	P	prima
offcie_r	R	připomínka (např. v týdnu po svátku světce se prošlý svátek připomíná některým z k němu náležejících zpěvů)
office_s	S	sexta
offcie_t	T	tercie
office_v	V	nešpory
office_v2	V2	druhé nešpory
office_x	X	není jasné, k čemu kus patří

A.2 Instrukce k instalaci nástroje

K připravení aplikace k prvnímu spuštění je potřeba následující:

- Mít stažené soubory pro běh aplikace. Ty jsou získatelné online[1] či z elektronické přílohy práce.
- Ujistit se, že je instalován Python (verze 3.11.6 a vyšší). Jestliže ne, pak pro instalaci postupujte podle návodu na oficiální stránce jazyka.[2]
- Nainstalovat jeho knihovny podle souboru `app_requirements.txt`, který je k dispozici ve složce `app`. (V příkazové řádce spustit `pip install -r path/to/app_requirements.txt`.)

Po úspěšné instalaci všech závislostí už stačí jen v příkazové řádce ve složce `app/Gregorian_chant_repertoire` spustit příkaz `python manage.py runserver`. Následně je nástroj k nalezení skrze webový prohlížeč na adrese `http://127.0.0.1:8000/`.

[1] `https://github.com/DvorakovaA/Mapping_the_Repertoire_of_Gregorian_Chant/tree/8b986a6ae841a4989f53e4a5dcf764c8a9ff07e6/thesis/app`

[2] `https://www.python.org/downloads/`