TECHNICAL REPORT

Issues in Cross-Database Harmonisation in the Cantus ecosystem

As identified during preparation of the CantusCorpus v1.0 dataset

Anonymous Authors

June 30th, 2025

During the process of preparing CantusCorpus v1.0, we identified numerous incompatibilities between individual Cantus network databases, both for chants and sources. These mostly stem from a lack of controlled vocabularies, which in turn leave opportunities for divergent editorial policies. This openness has been a feature that makes Cantus network attractive for individual database directors to join — the cost of joining is just implementing the Cantus ID mechanism rather than abandoning what are in some cases decades of editorial policy within a chant scholarship tradition. However, as we try to make full use of the aggregated data across the entire Cantus ecosystem, this lack of standardisation is now becoming an obstacle to computational research, already in basic steps such as aggregating chants by office. Answering even basic questions such as "Is Vespers repertoire from France different from that of Poland?" runs into two issues: Vespers are not necessarily marked with a consistent "V" in the `office` field, and provenance of sources is not necessarily given in ways that can be interpreted as French vs. Polish: one may easily encounter a record saying simply "Abbatia Lubensis", which without gives no indication to a computational filter that it refers to the monastery in Lubiaż, Upper Silesia (Poland).

This absence of controlled vocabularies or links to authority files is not something one can sustainably resolve by leapfrogging editorial decisions made at the level of Cantus Index board of directors and the directors of individual databases for the purposes of releasing a dataset — one may inadvertently gloss over real uncertainties. Instead of trying to standardise CantusCorpus v1.0 according to some arbitrary decisions made by us, we provide this document as:

1. A listing of main issues that dataset users should be aware of.
2. A roadmap for further standardisation across the Cantus ecosystem.

This standardisation effort is ongoing. Most importantly, the Cantus Database has completed and published its annotation manual,[1] which provides clear guidelines on how to use the different fields specified by the Cantus Index API. The Portuguese Early Music Database (PEM) has written guidelines as well.[2] These standards are not necessarily being applied in other databases, but at least for the Cantus Database and PEM we thus have a clear meaning of the individual values (such as modified mode). If we had similar guidelines available from other

[1] `https://cantusdatabase.org/documents/`
[2] `https://pemdatabase.eu/node/93860`

databases, it would have perhaps been possible to perform further harmonisation steps, but because these guidelines are not fully documented, it is not yet possible to map the different usages of individual fields between databases to each other.

What follows is our current understanding of what is and is not standardised across the Cantus Index network of databases: both for chant records, and for source records.

1 Chants: Genre

In scraped Cantus Index data, besides standard genre values as defined in the Cantus Index genre list,[3] there occur instances of genres such as:

- `In4`, `A14` - probably encoding position in liturgy
- `Ant/Resp`
- `Dox` - doxology
- `HymnV` - dealing with strophic texts, same as `SeqV`
- `a5+`
- `R+`
- etc.

Discarding the numbers or pluses etc. might mean something else – we do not know where these numbers belong, exactly.
This is the only field in which active harmonisation efforts were undertaken, with the Cantus Index interface and its genre list being instrumental in this endeavour. Because we obtain lists of existing Cantus IDs by listing them on Cantus Index by individual genre from this list, we can assign to the chant records this "I am from this CI genre list" value rather than the non-standard values which the individual database editors used. We are thus still working downstream of editorial decisions — decisions taken at the Cantus Index level rather than the component database level, but still decisions not taken by us.

2 Chants: Feast

The values of field feasts are subject to considerable variation. The Cantus Index lists 1794 feast names,[4] whereas the field feasts comprise 2401 different values in the available data. This situation is, as we believe, caused by a number of different scenarios not limited to the following suggestions.

Designations such as `H6/f5` or `4f3` coming from the Hungarian Chant Database mostly match onto Cantus Index feasts but are marked in this different marking system ("tempus" and "dies").

Names coming in sort of sequential manner (season, week, day), with instances such as `Adv., hebd 4., sabbato` (vs `Sabbato Hebd. 1 Adv.`) or `Annuntiatio Mariae, infra oct.` (vs `Annuntiatio Mariae,8`) are being recorded by PEM. The positive aspect of this situation is that, in principle, they can be matched with `feast_code` to Cantus Index names once the relevant data has been obtained and thus unified under one standard in the produced dataset.

Finally there are the feasts that are probably not present in the Cantus Index because nobody added them outside of some particular database such as `Odilonis` or `Rudesindus`. Number of those "not in CI list, not PEM, not HCD" feast values is approximately 320, where some of these are present in more than one source database. And

[3] `https://cantusindex.org/genre`
[4] `https://cantusindex.org/feasts`

we definitely miss some other scenarios bringing unmatched duplicates (e.g. `Ludmila` and `Ludmilla` representing one saint) or other kinds of standardisation issues.

It is important to note that not all feasts are accompanied by a `feast_code`. Even within the Cantus Index list there are 177 feasts for which the `feast_code` is empty. This observation indicates that `feast_code` as a principle of solving this is not currently a viable solution.

There is ongoing standardisation effort in the Cantus Database, also there is the LFRI initiative list[5] as well as Usuarium collecting the liturgical calendar.[6] We add the `feast_code` field based on the feast vocabulary on Cantus Index, but feast is a field that is undergoing changes as part of the standardisation process and is expected to change in future versions.

3 Chants: Office

When cleaning collected chant records, we looked into the variability of office annotations. Besides standard expected values (those from Cantus Database list) and Hungarian numeric values coming from "drupal issue", there were 11 more values (that became 17 after Hungarian numbers resolving since some of their textual abbreviations are also non-standard) present in the office field (e.g. `S&O`, `C2` or `CC`). The biggest question mark raised is whether `MASS` (used in SEMM and in a few other databases) and `MI` (proposed by CD) are the same office and though can be unified (see also `??`).

4 Chants: Mode

It may be surprising that overall 442 different values we found in the mode field. Beside numbers `1` to `8`, their transposed variants marked as `1T -- 8T`, their variants marked as `1S -- 8S` and `r`, `*`, `?` many different values exists in the data:

- textual representations (e.g. `Authentus tetrardus` or `protus`)
- numbers with ! (e.g. `4!` or also `-!` or `1!*`)
- different marks of transposition (e.g. `3 Trp` or `rtrp`)
- various question marks with numbers (e.g. `7T? ?T`)
- things like `IIII` , `yd = 7` , `FC`
- and quite a number of others.

5 Chants: Folio

Although it is not a troublemaker field, considering the narrow, but potentially important, use and given that it is a mandatory field, we feel the need to mention that even here are a few deviations from the otherwise fairly held standard. That includes mainly:

- foliations like `142v-143v` including not only the initial folio but the entire range of chant
- and missing leading zeros: `88v` instead of `088v`, that can break ordering of pages.

6 Sources: Century

While preparing code for the numeric century field assignment, we meet many variants of, more or less verbose, expressions of date of origin for manuscripts, e.g.:

- `12th century`

[5] `https://lfri.pemdatabase.eu/feasts`
[6] `https://usuarium.elte.hu/calendarlabels`

- `c. 1200`
- `late 15th century`
- `13th century (1275 - 1300)`
- `possibly around 1225-50`
- `mid 16th century [c.1540-c.1555]`
- etc.

These values are chosen well for display in the database front-ends, but for computational processing (already for e.g. visualisation on a timeline) they need to be standardised. This may have to include some indication of uncertainty.

If the value spans more than one century, the Cantus Index API description says to use the lowest number. This policy only exists because Cantus Index defines the century of the source as a (non-compulsory) property of a chant, though in practice this field is not used for chants at all.

7 Sources: Provenance

It is difficult to geolocate the places of origin of manuscripts and fragments not only because of their disappearance throughout history or a bit of forgetfulness, but also because of the high variability of names occurring for one place. We have 642 provenance field values within the 2278 sources collected, and this is also because of the general lack of a standard for coding uncertainty.
Among the examples we found are:

- `St-Martial` (CD) and `St. Martial` (CD) and `St Martial de Limoges` (MMMO)
- `Praha, klášter sv. Jiří` (FCB) and `Prague, St. George Monastery` (CDB) and `St George's Convent in Prague` (FCB)
- `Italie` (MMMO) and `Italy` (CDB)
- `Slovensko - Bratislava` (CSK) and `Bratislava` (CSK)
- `Sud de la France ou Espagne` and `Londres` and other values in French (MMMO)
- `See Description"" above""` (SEMM)
- `Central Europe / Slovakia / Ordo Cartusiensis` (CSK)
- etc.

It is evident that other fields such as `position`, where for example in Matins for the first antiphon of second Nocturns one can use position `2.1` or `7`, could give rise to a number of other problems.

We have so far chosen *not* to normalise these fields to a best-effort standard. This will have to come through subsequent efforts within the Cantus Index community.