A FRBR$_{OO}$-based annotation ontology for digital editing

Peter Boot, Huygens ING (KNAW)
Marijn Koolen, Humanities Cluster (KNAW)

Introduction

Digital editions of books and manuscripts are usually web sites, and often we refer to these editions as well as to their constituents by their URL. However, underlying the edition is a bibliographic universe which consists of the witnesses for the edition (physical objects and their parts) as well as the textual or conceptual objects (works and (printed) editions) for which the physical objects are the bearers. Beyond this pre-existing collection of objects, work on the edition creates a new set of objects that represent the physical and logical bibliographical objects: digital images, transcriptions (typically XML), and renderings (HTML pages integrating digital images, transcriptions and further editorial additions, such as metadata, notes and translations).

This paper will present an ontology for describing the content of digital editions, both the objects (documents and works) that were edited and the outcome of that edition process (digital text and images). Such an ontology is necessary for interoperability between the digital edition and the wider world of digital scholarship. The use case that the paper discusses is scholarly annotation: if we want our annotations to be reusable we need to be sure that they are attached to the proper object: they shouldn't target, say, Goethe's *Faust* while they are actually about one line in a particular manuscript; they shouldn't target a specific image if the annotation is really about the page that the image represents.

Boot et al. (2017) argued that to support scholarly annotation, the editions' HTML should be extended with pointers to the underlying data structure. They discussed only briefly the question to what model that underlying data structure should conform, providing an ad hoc ontology. This paper will present an ontology that is based on FRBR$_{OO}$ (IFLA 2015)[1] and builds heavily on the DET scheme developed by Peter Robinson and his colleagues (Robinson 2017). It takes into account the multiple levels present in the edition: the conceptual work, the witness(es) to that work and the resulting output. We present our ontology as a proposal for discussion. At the conference, we will give an example based on an edition of the letters of Vincent van Gogh.

In order to keep the discussion manageable, we ignore for now the complications arising from complex works (Container works, in FRBR parlance, that consist in the arrangement of individual other works), and all forms of editorial enrichment (such as annotations, translations and apparatus). We also ignore works with important visual components.

[1] Including the overarching architecture of the CIDOC/CRM model (ICOM/CIDOC CRM SIG 2017).

Modelling proposal

Scholarly annotation can only be successful if the user can indicate clearly what the target of the annotation is. This apparently simple requirement is in truth hard to satisfy. A user may want to annotate the text of the edition (e.g to point out a transcription error), the text of the edited document (say to draw attention to the peculiarities of the scribe's writing) or the text of the work (e.g. to explain a historical reference). An edition may also contain multiple text representations, e.g. diplomatic transcriptions representing witnesses and a reading text representing the work.

A very general representation of the editable (bibliographic) domain might look like figure 1. Editing proceeds from (material) documents, that have a double nature: they are both physical objects that consist of physical parts (the text bearers) and they contain a text (that can have parts). The text is a manifestation of a more abstract work that may also have manifestations elsewhere. A Positioned Text Fragment (PTF) is a sequence of signs belonging to the text as given in a certain physical location; they are the things that Peter Robinson's (2017) DET scheme was designed to identify (and which he calls 'Text' or 'TextFragment').[2]

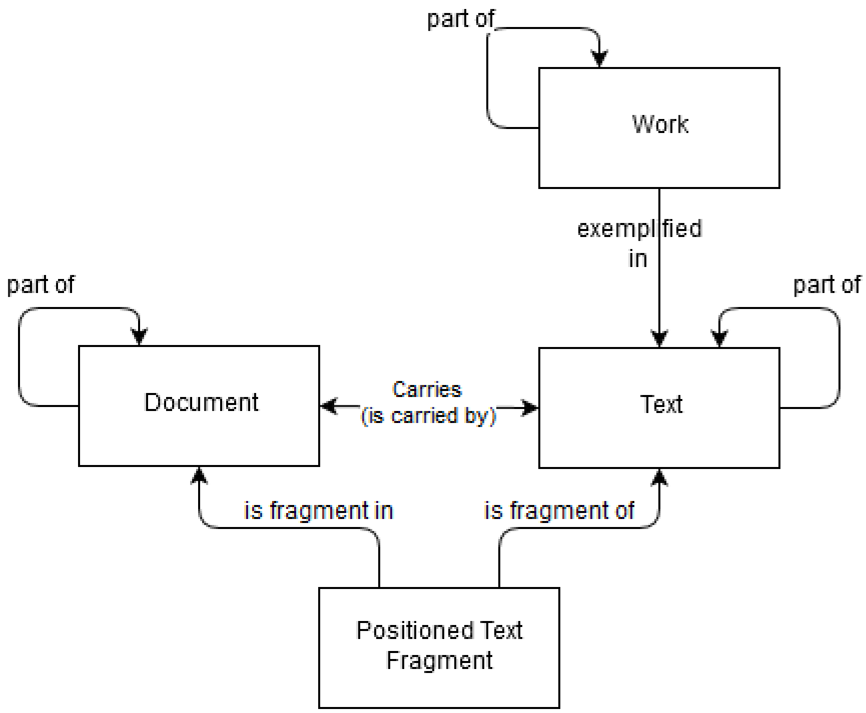

Figure 1. The editable domain

These entities and their relations correspond to objects and properties in the $FRBR_{OO}$ model (IFLA 2015). A document is typically a manuscript or a copy of a book. In $FRBR_{OO}$ terms, an author's manuscript is an F4 Manifestation singleton; a copy of a book is an F5 Item. As a physical object it can consist of parts (P46 is composed of; a property from CIDOC/CRM). The

[2] Robinson prefers 'Entity' over 'Work', but in an information modelling view, anything can be an entity.

text that the document carries is an F2 expression; in the case of an authorial manuscript more specifically an F22 self-contained expression, in the case of a printed book an F24 Publication expression. FRBR does not describe the relation between works or expressions and their parts (say a poem and its stanzas). The fragments are F23 Expression Fragments. The result is given in Figure 2 (FRBR$_{OO}$ terminology in red).[3]

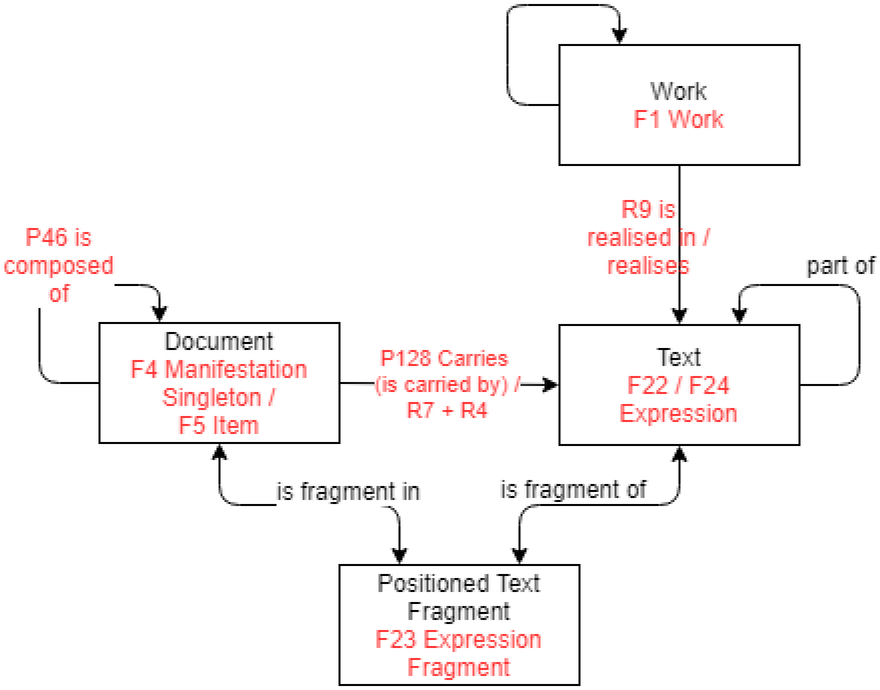

Figure 2. Editable domain, with references to corresponding FRBROO concepts

We move on from the domain of the editable to the domain of the edition. The edition can provide (and/or): page images, an edited text close to the manuscript version, or a text intended to represent the work. As a publication, the edition is itself part of the domain that is covered by the FRBR$_{OO}$ ontology. Figure 3 gives a simple representation, ignoring for now the distinction between the visible components of the edition and the underlying technical objects (e.g. XML and image files).

[3] The relation between document and text is somewhat convoluted in FRBR$_{OO}$: an F4 Manifestation Singleton (manuscript) *P128 Carries* the expression; however, an F5 Item (printed book) *R7 is an example of* an F3 Manifestation Product Type, which *R4 Comprises carriers of* the expression.

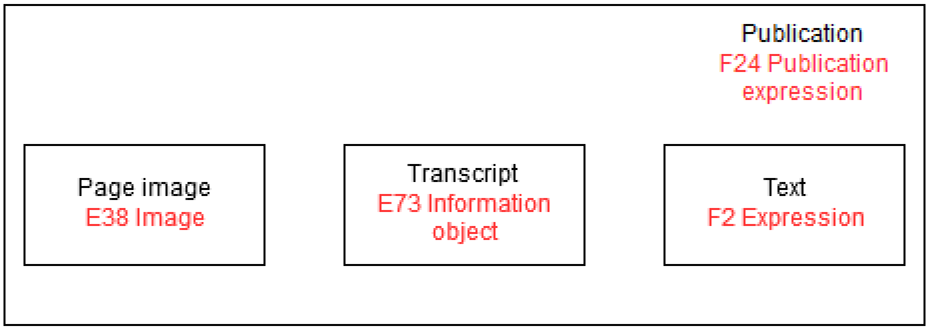

Figure 3. The edition

The edition's components represent the objects in the editable domain, as shown in figure 4. In this representation we leave out the Expression level, as that can be represented as a (the largest) Positioned Text Fragment (and it simplifies the diagram). The text bearers in the physical realm are represented by page images (in CIDOC/CRM: P138 represents). The transcript represents the PTFs (P138 is technically limited to images but seems appropriate here as well). The work may be represented (P9 is realised in) if the edition provides a text that goes beyond a transcription of individual witnesses. The edition is also a publication expression of the same work that the documents express.[4]

[4] Technically, in FRBR_{OO} terms, the edition as whole is an F19 publication work. There is no direct relation between the authorial work and the publication work, except that the F24 Publication expression *P165 incorporates* the expression of the authorial work. We have labeled the resulting indirect relation as 'Member of'.

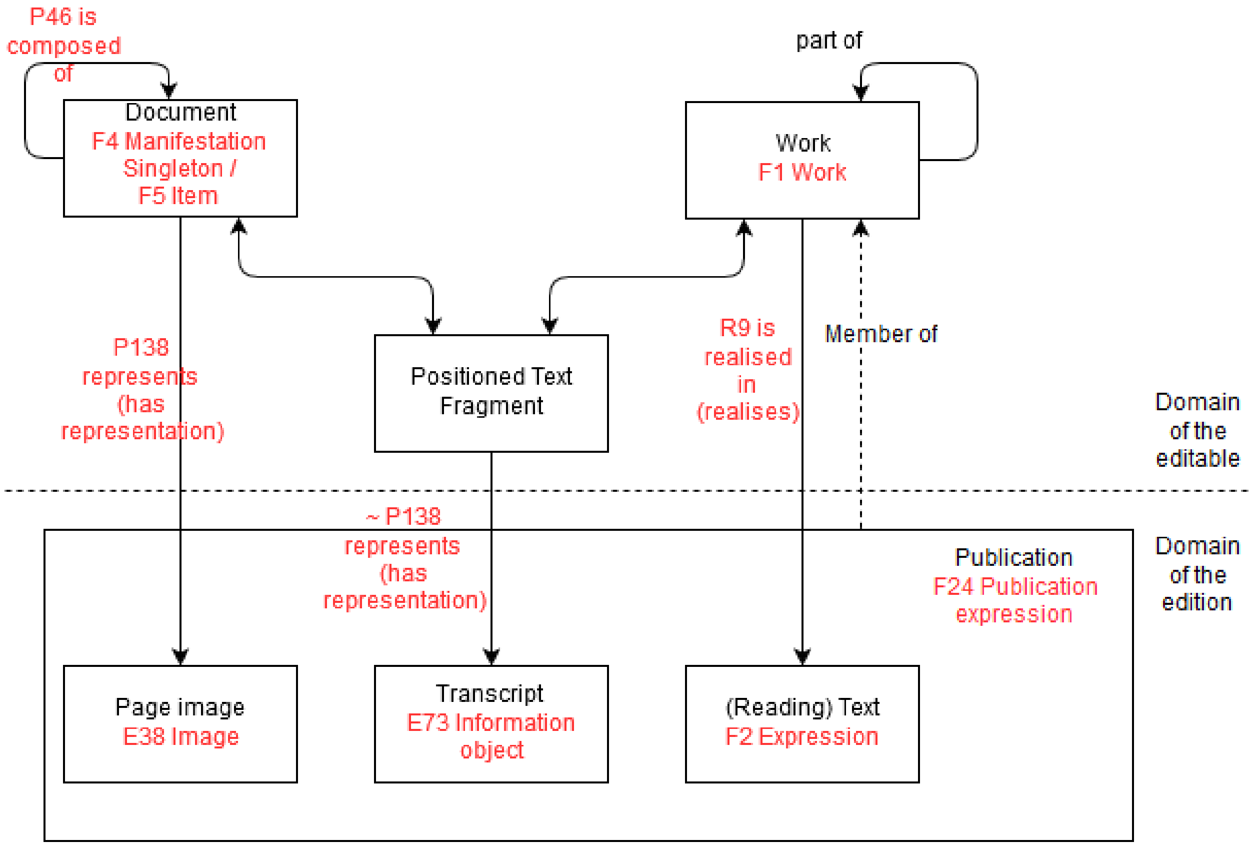

Figure 4. The edition as a representation of the bibliographic domain

This finally puts us in a position to show the top of our annotation ontology (Figure 5). All annotation to the edition (except for annotations like 'zoom function doesn't work' or 'font too small') is either about the bibliographical realm that the edition represents, or about the claims that the edition makes with respect to that realm, i.e. about EditableThings or EditionThings. When we annotate EditableThings, we make a claim about the outside world, e.g. a manuscript; when we annotate an EditionThing, we make a claim about how that external object is represented in the edition. The subclasses of EditableThing and EditionThing correspond to documents, works and the positioned text fragments that represent (parts of) works rendered on (parts of) documents. Technically, we define (parts of) documents as the union of documents, physical parts of documents and zones on those parts (F9 Places).

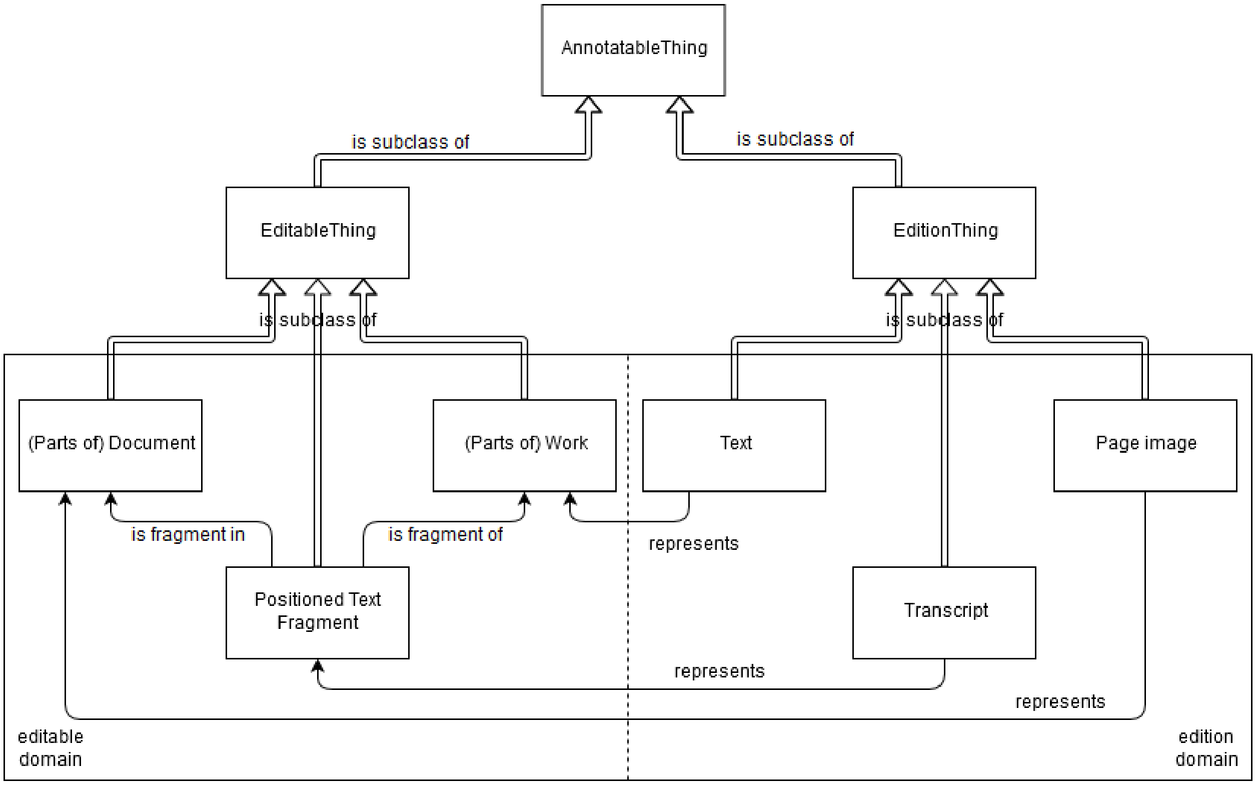

Figure 5. Top ontology for annotatable things in the domain of the scholarly edition.

Conclusion

We have shown a FRBR-based model of the domain of digital editing, both of the objects to be edited and of the editing's output. In many aspects it is comparable to the model underlying Peter Robinson's Textual Communities project.[5] The model provides the basis for an ontology for the domain of scholarly annotation in the digital edition. Next steps include extension of the model to cover editorial enrichments and the situation of container works. What also needs elucidation is how to integrate the relevant information into TEI/XML source files. Discussion of the integration between TEI and ontologies has up to know mostly focussed on events, persons and tangible objects (Eide 2014) and on an ontological version of the TEI abstract model (Ciotti 2018); integration of the core objects of editing in the LOD world is to the best of our knowledge as yet largely unexplored.

Apart from these technical challenges, to which an answer can surely be found, a perhaps more difficult problem may be how to explain to the user of the edition that there are differences between annotating the page image and the page or the transcript and the positioned text fragments. A carefully designed user interface will be essential to avoid confusion; we are

[5] http://www.textualcommunities.usask.ca/web/textual-community.

experimenting with that in the context of a medieval miscellany. Still, a sound underlying model is essential for interoperable annotation in the digital edition.

References

Boot, P., Haentjens Dekker, R., Koolen, M., & Melgar, L. (2017). Facilitating Fine-grained Open Annotations of Scholarly Sources. *Digital Humanities 2017*. https://dh2017.adho.org/abstracts/198/198.pdf

Ciotti, F. (2018). A Formal Ontology for the Text Encoding Initiative. *Umanistica Digitale* 3.

Eide, Ø. (2014). Ontologies, data modeling, and TEI. *Journal of the Text Encoding Initiative*, (8).

ICOM/CIDOC CRM SIG. (2017). *Definition of the CIDOC conceptual reference model*. Version 6.2.2. http://www.cidoc-crm.org/sites/default/files/2017-09-30%23CIDOC%20CRM_v6.2.2_esIP.pdf

IFLA. (2015). *Definition of FRBR$_{OO}$: a Conceptual Model for Bibliographic Information in Object-Oriented Formalism*. https://www.ifla.org/publications/node/11240.

Robinson, Peter M. W. (2013) “Towards A Theory of Digital Editions”, *Variants*, 10, pp. 105-132.

Robinson, P. M. W. (2017). Some principles for making collaborative scholarly editions in digital form. *Digital Humanities Quarterly*, 11(2).