

Report

Entity-fishing for Scholarly Publishing: Challenges and Recommendations

Andrea Bertino, Luca Foppiano and Javier Arias, Aysa Ekanger, Klaus Thoden

On 4th September 2018 the [Göttingen State and University Library](#), with the support of the [Max Weber Stiftung](#), organised the second HIRMEOS Workshop on [Entity-Fishing for Digital Humanities and Scholarly Publishing](#).

Entity-fishing, a service developed by [Inria](#) with the support of [DARIAH-EU](#) and hosted at [HUMA-NUM](#), enables identification and resolution of entities: named entities like person-name, location, organizations, as well as specialist and less commonly classified ones, such as concepts, artifacts, etc. The technical specifications of the service are described in the paper "[entity-fishing: a service in the DARIAH infrastructure](#)", while for a quick overview of how it has been implemented on the digital platforms involved in the HIRMEOS project you can have a look at this [factsheet](#) and at the recording of the [webinar](#) organized by the SUB Göttingen.

The workshop aimed to discuss and clarify practical concerns arising when using the service and possible new use cases presented by [Edition Open Access](#), [ScholarLed](#) and [Septentrio Academic Publishing](#).

This report describes challenges related to the development of these applications and provides recommendations for its integration and use on digital publishing platforms. The solutions proposed can be further applied on other domains.

Beyond the *tradigital* format: the need for TEI XML

The new consortium ScholarLed is developing a common catalogue of all monographs published by several scholarly presses, with a recommendation system that will suggest similar books to those browsed by the user. While the initial proposal expected the similarity linkage to be done manually, ScholarLed intends to explore if *entity-fishing* could automate the process, finding common entities within the catalogued books. Furthermore, ScholarLed is exploring other possible use cases within the context of improving the discoverability of their publications.

Indeed, there is great potential in adopting wikidata entities as a standard metadata export. What still needs to be clarified – observes **Javier Arias** from [Open Book Publishers](#) – is how this data is to be disseminated and promoted to distribution platforms and data miners. In order to increase the interest in this service, the presses involved in ScholarLed think that it is essential to find the best way to export the annotations and associated Wikidata IDs along with other book metadata, avoiding these data getting lost during book redistribution or reprocessing. To this end, it would be important to embed the found wikidata entities in their [TEI XML](#) files – a feature that at present is not yet available via *entity-fishing* service.

OpenBook
Publishers 


ScholarLed

ScholarLed: Members

OpenBook
Publishers 

 MATTERING PRESS

 meson press

 punctum books

 OPEN HUMANITIES PRESS

 may fly

With regard to the realisation of a shared catalogue for ScholarLed, Javier Arias pointed out in his [presentation](#) some specific challenges concerning its implementation in the platforms, the accuracy of the service and the dissemination of entities and metadata.

In a similar way, **Klaus Thoden**, who has tested how to enrich the content and extend the usability and interoperability of the books published by *Edition Open Access*, argues that it would be extremely important to add some kind of support for TEI XML files, allowing the user to input an unannotated XML file and get back the file with the entities embedded as TEI annotations. In the case of Edition Open Access a PDF file is just one output format amongst others and only supporting TEI XML would make the publications annotated through *entity-fishing* fully reusable.

Usage Scenarios

The screenshot displays a digital edition interface. At the top, a map shows a geographical area with a red line indicating a path. Below the map is a search results panel titled "Keywords, Persons and Locations". This panel lists several entries with associated page numbers: Francesco I.: 1; Ferdinando I.: 1 2; Vestrucci, Lorenzo: 1; Florenz: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89; Siena: 1 2 3 4 5 6 7 8 9 10; Bologna: 1 2 3 4 5 6 7 8 9 10 11; Nipozzano: 1; Raticosa: 1 2; Filigare: 1 2; Firenzuola: 1; Scaricalasino: 1; Romagna: 1 2. To the right of this list, a section titled "Also appears in:" shows two book covers: "The Civilization of Knowledge in the Medieval World" and "Censibile". Below the search results, there is a snippet of text in German: "vor, ein schnadnates Straßenspiaster per mano di buono maestro reparieren zu lassen. Für Straßenspiasterungen, die auf Hauptstraßen die Regel waren, wurden in der Regel Handwerker bezahlt. Tabelle 2.1 zeigt – gegliedert nach Regionen – die Kosten für Straßenspiasterbau und -instandhaltung während der Regentschaft Ferdinands I (1587–1608)." At the bottom of the screenshot, the URL "http://edition-open-access.de/studies/5/4/index.html#78" is visible.

Four usage scenarios for Edition Open Access in [Klaus Thoden's presentation](#): cross-publication discoverability of entities; interconnection of publications; links to taxonomies; search and browse functionality for entities

There are therefore some inherent difficulties with the service in relation to the use of the PDF format, which, although allowing the user an experience of the digital document as similar as possible to that of a traditional printed document, seems to somewhat limit the reuse of the annotated publications. See Luca Foppiano's [technical assessment](#) of how to process input in TEI XML format.

Luca Foppiano notes that it could be interesting to allow the manual annotations of PDFs, however the resulting feature might not perform as well as expected and implementing might not be worth the effort.

The problem rests in how the data is segmented internally. By referring to the [GROBID library](#), the PDF is divided into a list of *LayoutTokens*: every single word corresponds to roughly one token but this can vary depending on how the PDF has been generated. A *LayoutToken* also contains coordinates, fonts and other useful information. It is important to understand that the process is not always working perfectly. The order of the tokens might not be corresponding to the real order, and there are other small problems.

As of today, *entity-fishing* is the only tool available that maps entities from the coordinates extracted from the PDF. However, the opposite direction is more challenging as the coordinates

associated with an entity might correspond to one or more layout token, thus matching them might give imprecise results.

A solution to this challenge could be the development of a new graphical user interface to load a PDF and give the user the possibility to manually annotate it. The output annotations will be then compatible with the *LayoutToken* detected coordinates.

How much editorial work is required to validate the outputs of *entity-fishing*?

Scholarly publishers willing to adopt *entity-fishing* in order to allow deeper interaction with their content are interested to maintain a balance between the benefits of such implementations and the amount of editorial work associated with them. It is necessary to develop a workflow that allows a smooth curation of the data obtained from the system, but it is difficult to assess how much time is required for the curation of the extracted entities without the editors having the system tested. Therefore, the first step in implementing *entity-fishing* must be to clarify carefully the specific kind of service the publisher is looking for, rather than discovering possible applications on the go. Klaus Thoden notes that since *entity-fishing* is a machine driven approach and can lead to results that are clearly wrong or wrongly disambiguated, the user must also be made aware of how the results came about, as the credibility of a source suffers from wrong results. If, however, it is made clear that the displayed enrichments result from automatic marking, the user records the information with a pinch of salt.

How to store data for the extracted entities and present full-annotated text?

By using *entity-fishing* to enrich the content of digital publications, we need to store the data for the extracted terms. At least two solutions are possible. A first possible scenario is that the editor modify the contents of the paragraph fields in the database and add additional span elements. Otherwise, the data can be stored elsewhere by using standoff annotation per paragraph.

In addition, the viewer used to present the processed document should be adapted appropriately. For example, to separate the authored content and the automatic tagging, the publication viewer should contain a button where automated enrichments can be toggled. Thus, the user still has a choice of seeing or ignoring these enrichments. Furthermore, a functionality could be implemented that allows correction or flagging of incorrect terms.

Can *entity-fishing* be used to disambiguate digitized old documents?

Aysa Ekanger was exploring the possibility of using *entity-fishing* in order to recognize old toponyms and person names, and automatically disambiguate them – in a series of digitized books, [Aurorae Borealis Studia Classica](#), and possibly in some historical maps that are now being digitized at the University Library of Tromsø. Old maps present an interest both to scholars and to the general public. For digitized maps, two hypothetical use cases were discussed: annotations and using *entity-fishing* to produce interactive map-viewing, with an old map and a

modern equivalent side-by-side.

Another instance of map production can come from [Septentrio Reports](#), where a lot of archaeological excavation reports from Tromsø Museum are about to be published (this is born-digital text). Apart from toponym entries in annotated PDFs, the *entity-fishing* tool could be used to produce a map of excavations. The map could be visualized through an appropriate tool and placed on the level of an issue, as one issue should contain 6–7 reports from different excavation sites.

Old digitized texts present some specific challenges. For instance, there is a custom of referring to scholars by latinized names: Johann Christoph Sturm is referred to as “Sturmius” in [this](#) German text from 1728. Old toponyms that would be disambiguated would be of at least two types: names of places that no longer exist and archaic names of existing places (including old orthography). In order to achieve this kind of disambiguation there are two requirements: 1) the terms are contained in a knowledge base, and 2) there is a piece of code which can recognise the relevant tokens in the text.

darüber: und je näher es dem Tropico kömmt, je schwächer bleibt es in seinem Schein.

Phenom. IX.

Je höher man in Norden kömmt, je klärer wird das Nord-Licht. Ich habe selbst darvon einen grossen Unterschied gemercket in Norwegen zu **Christiania** im 60. Grad und 28. Minuten, und zu **Drontheim** im 64. Grad, und können verschiedene Nordische Reise-Beschreibungen uns lehren, wie heller es da zwischen dem 70. und 80. Grad Latit. Boreal. erscheint.

Phenom. X.

Der gemeine Mann machet sich viele Praeligi und Wunderzeichen vom Nord-Lichte und deutet es doch wenigstens auf die Veränderung des Wetters und der Jahreszeiten: Ich habe aber nach zweijährigen fleißig angestellten Observationen gefunden wie falsch und ungewiß es sey: so daß bey der stärksten seltensten und merckwürdigsten Erscheinung des Nord-Lichts keine andere Folge bemerken können, als zu einer Zeit Kälte und Frost, zur andern Regen oder trocken Wetter, einmahl Sturm und Wind und das anderemahl stilles Wetter, ohne eine gewisse Ordnung und reguläre Folge.


§. II.

Archaic place names

Oslo

Cond. prob.: 019373d40752351097

(, norwegisch, oder) ist die Hauptstadt des Königreichs Norwegen. Ihr ehemaliger Name war **Christiaania** (1624 bis 1824) bzw. **Kristiania** (alternative Schreibweise von 1877/1897 bis 1924). Die Kommune Oslo hat 658.390 Einwohner (Stand: 1. Januar 2016). Sie bildet eine eigenständige Provinz (Fylke) und ist zudem Verwaltungssitz für die benachbarte Provinz Akershus.



DRONTHEIM

Normalized: Trondheim

Domains: Astronomy, Administration, Biology, Geography, Sociology.

conf. 0.8778

Trondheim, oder früher 'Dronthjem' geschrieben, deutsch veraltet: Drontheim) liegt an der Mündung des Flosses Nidelva in der Provinz (Fylke) Ser-Trondelag in Norwegen und wurde 997 als Nidaros gegründet. Trondheim ist mit Einwohnern (Stand) nach Oslo und Bergen die drittgrößte Kommune des Landes. Mit einer Gesamtfläche von 342 Quadratkilometern umfasst sie neben dem Stadtgebiet seit 1964 die umliegenden Siedlungen.

category for people who die here: Q9220220

<https://septentrio.uit.no/index.php/aurora/article/view/4302>

[Aysa Ekanqer's presentation](#) with three possible usage scenarios: PDF annotation, word clouds and digitized maps

A further relevant technical concern, however not directly connected to *entity-fishing service*, is the inaccuracy of the OCR: in digitized historic texts 1–8% of OCR results are wrong. Of course, the improvement of OCR is necessary, but in the meantime the editors of the text in question may have to resort to “manual” labour for a small percentage of the text: any text mining application would therefore require a preliminary (manual or semi-automatic) correction of the errors in PDF due to the OCR process.

Best practices on how to use *entity-fishing* in scholarly publishing

During the discussion, it became evident how important it is to have an **exact preliminary definition of the usage scenario** that a content provider such as a publisher intends to develop on the basis of *entity-fishing*. In particular, it is important to ascertain **how much work is required** to manage the specific publishing service based on *entity-fishing*.

From a more technical point of view, these are the key recommendations made by the developers of *entity-fishing*:

1. **Consider the web interface as a prototype and not as a production-ready application**

The demo interface shows “what can be done”. There is a step in the middle that is to adapt the service to your own requirements and this has to be done by a data scientist, someone that can quickly look up the data and get some information out of it. For example, the language R can be used to read, visualise and manipulate a list of JSON objects. In addition, you should avoid to evaluate the service on the basis of restricted manual tests. Proper assessment must be carried out with the correct tooling and should converge in a prototype showing some (restricted) results.

2. **Do not perform on-the-fly computation but store your annotations on a database to retrieve them when needed**

There are plenty of reasons to support this recommendation, the most important are:

- a. Usability, on-the-fly computation is overloading the browser, making it slower and unresponsive.
- b. Processing a PDF will take more than 1 second, if this delay is propagated to the UX, the user will think the interface is slow.
- c. For small paragraphs it can be done, but it has to be handled asynchronously in the interface or the user will notice.

Future projects and prospects

Indexing a corpus of publications by extracting keywords via *entity-fishing* can significantly increase the discoverability of these publications. During the workshop, however, we found that such indexing only makes sense if it includes a very high number of publications. In view of this, we can imagine forms of a **unified catalogue and a federated search engine** for publications of different platforms indexed through *entity-fishing*. The realization of such a catalogue seems to be something that goes far beyond the possibilities of a single institution and requires the action of distributed research infrastructures that, as in the case of [OPERAS](#), can coordinate the activities of several stakeholders. More precisely, OPERAS intends to use the *entity-fishing* service as one of the key components of its future discovery platform to be developed between 2019 and 2022.