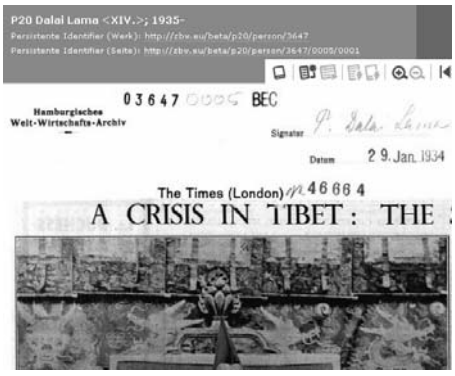
The 20th Century Press Archives as Linked Data Application

Joachim Neubert

German National Library of Economics (ZBW) – Leibniz Centre for Economics
Neuer Jungfernstieg 21, 20347 Hamburg, Germany
j.neubert@zbw.eu

The 20th Century Press Archives of the German National Library of Economics (ZBW) are a large collection of newspaper clippings about persons, companies, products and general subjects of public interest , covering the period from 1826 to 2005 and organized in thematic dossiers. Through our P20 application, the digitized page images have now been published as Linked Data, in order to give each item a persistent identifier for citation and linking. OAI-ORE provides the backbone for organizing the large and deeply nested aggregations of web resources. The aggregations are described by resource maps in RDFa and enriched by metadata from and links into the Linked Data cloud (especially DBpedia, the German Authorities Files, VIAF and Chronicling America). In the course of the project, both OAI-ORE and RDFa proved to be well suited for the organisation and presentation of archival collections with large amounts of digitized paper documents and few machine readable metadata.

Introduction



At the beginning of the 20th century, the Central Office of the Hamburg Colonial Institute[1] and the Economic Archives of the Kiel Institute for the World Economy started collecting information about politics and the international economy. This resulted in large Biographical, Companies, Products and General Subjects Archives covering almost every issue under public discussion. More than 1300 newspapers from all over the world were used as source material. For the Companies archives, annual reports and other material was added to the dossiers. The total number of documents is estimated at 30 million. Now held by the ZBW, these collections form a unique source for studies of German and international (economic) history of the 20th century from a German perspective.

[1] Renamed as Hamburg Institute of International Economics (Hamburgisches Weltwirtschafts-archiv - HWWA) in 1919

The documents until 1949 were digitized during a project funded by the German Research Foundation (DFG)[2]. The digitised material, sourced mostly from roll film, is being gradually prepared for publication [1].

During the digitization process, the thema of each dossier is captured in a label, which was copied from the physical boxes. For the Biographical dossiers, additional metadata – e.g. lifespan, profession and general biographical information about the person (in German), and especially the German Person Authority File Number (PND) – has been added intellectually. For the documents, additional information (such as title, author, source and publication date) has been added for 20,000 clippings during the digitization process. This labor-intensive work surely isn't affordable for the whole amount of 5.7 million already digitized documents, so metadata will remain sparse.

Goals in Application Development

The P20 application described here[3] provides online access to 6,800 dossiers and 250,000 documents relating to persons and companies.[4] Currently in a beta release, the application amends a prior web application[5] resulting from the above-mentioned project.

The main goals in the development of the new application were:

1. To give each and every collection, dossier, document, page and even search result set a persistent identifier for citation and linking, in order to support scholary use.
2. To provide context from metadata on the web to users and link to other data sources relevant to the domain, in order to enhance user experience.
3. To support the use of a standard image and metadata viewer (based on METS/MODS) familiar to users.
4. To facilitate automated harvesting of the data for service providers and encourage re-use of the data, for instance in cultural heritage applications such as Europeana or virtual research environments for historians.

General Design Decisions

As we wanted to achieve optimal accessibility and reusability of the archives' data, and also wanted to pull in data from external Linked Open Data sources (using the authority file IDs as an entry point to the data net), we decided early to adhere to Linked Data Principles.

The data at hand came with relatively few metadata; the basic information lies in the structure of the collections based on thematic dossiers, and in the sequence of documents within the dossiers (sorted chronologically by publication date).

---

[2] Project summary (in German): http://www.kulturerbe-digital.de/en/projekte/9_38_383332.php

[3] **http://zbw.eu/beta/p20**

[4] Other material which is not yet digitally available or subject to intellectual property rights restrictions, can be used for academic purposes on the premises of the ZBW in microform.

[5] http://webopac0.hwwa.de/digiview/

OAI-ORE as the Backbone of the Data Model

OAI-ORE seemed well suited for this kind of quite generic aggregations. It "defines standards for the description and exchange of aggregations of Web resources", which "may combine distributed resources with multiple media types" [2]. The `ore:Aggregations` (as "non-information resources") are described by `ore:ResourceMaps` ("information resources"). These aggregations may be nested arbitrarily deep, which was also a requirement. Resources may be part of multiple aggregations, and `ore:Proxy`s may be defined to reference aggregated resources in special contexts, such as the order within a dossier or the score within an result set. Altogether, *Goal 1* (persistent identifiers especially for aggregated press material and dynamically generated result sets) was achieved easily through OAI-ORE.

The ORE-OAI vocabulary was also developed with data exchange and effective harvesting of large collections in mind (e.g. for every ResourceMap a `dcterms:modified` property is required). It exactly defines the borders of each Aggregation, distinguishing between aggregated resources (including multimedia files) and arbitrary internal and external web resources linked to elsewhere Fig. 1 gives a high level overview over the P20 data model. *Goal 4* (to facilitate harvesting) was also achieved by this model.

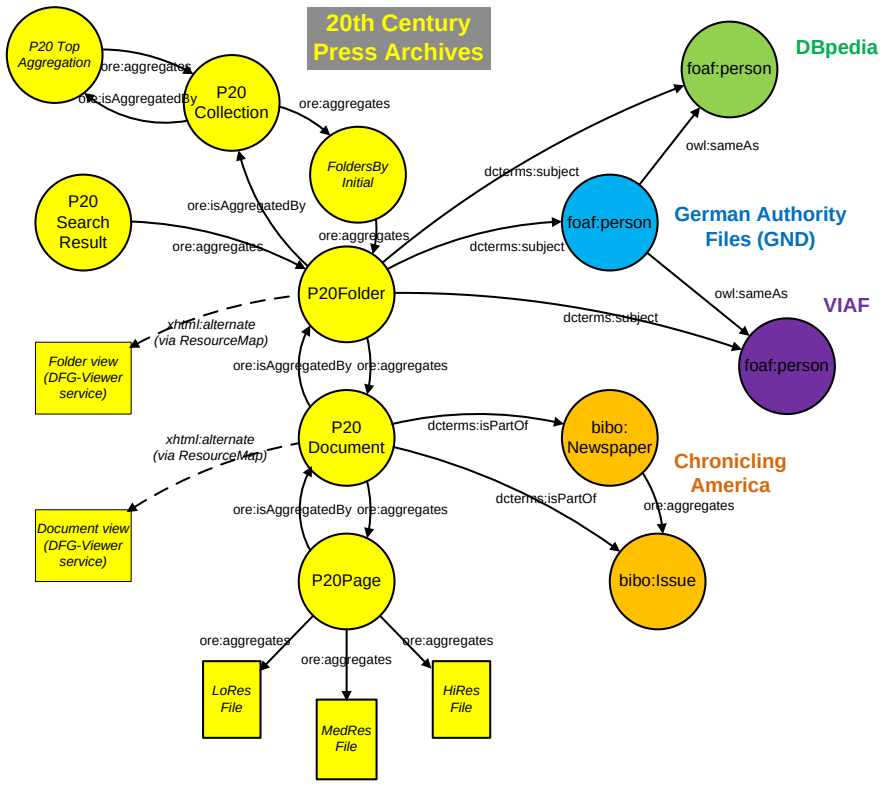


Fig. 1 RDF entities and other web resources of P20 and links into the LOD cloud

RDFa for a Single Presentation to Humans and Machines

Since the application was mainly aimed at end users, RDFa[6] was a natural choice for serializing ResourceMaps of Aggregations by embedding the data in (X)HTML pages. Thus, user navigation paths through the pages of the application are rather parallel to the RDF structure shown above.

As an additional advantage, much of the experience gained, and even program code, from the former LOD publication of STW Thesaurus for Economics [3] in RDFa could be re-used during development.

URI Concept

The URIs in P20 fulfill two central purposes. They work as

1. persistent identifiers, dereferenceable through HTTP, according to linked data principles
2. user interface as well as API of P20 as a web service (building on REST architectural style)

The `p20` prefix currently maps to `http://zbw.eu/beta/p20/`, in order to indicate the current beta state and the fact that URIs will change in the released version of the application. We plan to deploy a redirect mechanism for the URIs made obsolete.

Aggregation URIs:

```
p20:{collection_name}(/{dossier_key}(/{document_number}(/{page_number})?)?)?
```

Aggregation URIs are 303-redirected to

Resource Map URIs:

```
p20:{aggregation_uri_part}/about(.{language}.({format})?)?
```

With the generic `about` URI, language and format are content negotiated by Apache. Users of the data are however allowed to choose the representation explicitly. The application itself does this to make language switches by the user sticky. Supported languages are German (`de`) and English (`en`). Supported output formats are currently RDFa (`html`) and – for dossier and document – METS/MODS (`xml`).

View URIs:

```
p20:{collection_name}/{dossier_key}(/document_number})?/view(.{language})?
```

View URIs are mapped to a call to the DFG-Viewer[7] web service with the URI of the METS/MODS XML representation of the dossier or document as a parameter.

Search URIs:

Search result sets should be referenceable by URIs, too. The aggregation URIs

```
p20:{collection_name}/searchresult(/{language})?\?q={query}
```

are 303-redirected to

```
p20:{collection_name}/searchresult/about(/{language})?\?q={query}
```

This is currently implemented for Company dossiers (searching the title as free text).

---

[6] http://www.w3.org/TR/rdfa-syntax/. RDFa, RDF/XML and Atom are defined as serialization syntax for OAI-ORE

[7] http://dfg-viewer.de/en

Details

This section highlights the exploitation of Linked Data features for user interface enrichment and retrieval (thus achieving the above stated *Goal 2*) as well as implementation details and open issues.

Taking Advantage of the Linked Data Cloud
Since the metadata for Biographical dossiers already included PND id numbers, and since the German National Library had recently published the German Personal Name Authority Files as Linked Data[8], this provided an entry point to the Linked Data cloud. PND entries are mapped via `owl:sameAs` to VIAF[9] (and thus to supplementary bibliographic resources all over the world) and to DBpedia as the main Linked Data hub. Abstracts and Thumbnails from DBpedia are loaded into the page and supplemented with links to the corresponding Wikipedia articles. It should be emphasized that this approach allowed us to complete the English version of the application (which could initially only present data in German) at almost no cost.

Although document-related data is sparse and only a fraction is sourced from American newspapers prior to 1922, the links to Chronicling America showcase the potential of linking to newspaper sources available as Linked Data.

Linked Data enhanced Biographical Dossier Search
The Biographical dossier search function, which works as an autosuggest lookup on personal names, uses linked data from the above-mentioned name authority file, mediated through a web service[10], to add alternate names to the list of suggestions and to forward to the actual dossier:

```
Netto, Henrique M. → Coelho Neto, Henrique M.; 1864-1934
Reiling, Netty → Seghers, Anna; 1900-1983
```

Generally, the use of authority files and thesauri enhances users' search experience largely by offering access through synonyms which may be syntactically completely unrelated, but refer to the same person or same concept. The LOD publication of this type of data opens a way for arbitrary applications to enhance their retrieval functions without own investment in terminology development.

Search Results: Extending OAI-ORE to Dynamic Aggregations
To my best knowledge, ORE is applied here for the first time to search result sets, which are handled as dynamically built aggregations. It is used for Company dossiers[11], where generally only the dossier label is known. A search result set like `p20:company/searchresult?q=hamburg` can be referenced by URI. The main advantage however lies in the fact that it can be harvested – for example by a local history portal

[8] http://www.d-nb.de/eng/hilfe/service/linked_data_service.htm.
[9] http://www.viaf.org/
[10] http.//zbw.eu/beta/stw-ws, which in turn queries a SPARQL backend.
[11] Document Search was not implemented yet. Due to the relatively sparse metadata, results of a document search could be misleading.

– with all its dependent entities and files through standardized OAI-ORE harvesting, just like any other ore:Aggregation.

METS/MODS Mapping for Consumption in DFG-Viewer

In order to comply with a national quasi-standard (and to use a well working Open Source image viewer which was already available), the DFG-Viewer was chosen for viewing and browsing documents. It processes METS/MODS XML files for documents or whole dossiers. Since up to now no general mapping methodology from OAI-ORE to METS exists [4], since multiple levels of aggregations had to be mapped to a single METS file, and since the application profile for the viewer added some special conventions, we ended up with a custom mapping directly from the database structures. The persistent identifiers are offered from within this application, too. This allows users to work in a familiar environment and achieves *Goal 3* stated above.

Subdividing Large Aggregations

On the dossier level, aggregations are too large for intellectually browsing web pages as well as for efficient harvesting. Since a paging mechanism for resource maps [5] is still lacking (which is also a problem for large search result sets), we introduced an intermediate level, hashed by initials (see Fig. 1). The structure reflects the intent to provide "logical" navigation paths without the burden of having to parse or crawl large amounts of links (therefore we deliberately didn't create all inverse links here).

Technical Implementation

The development was carried out in a Perl environment; the application runs on an Apache web server under mod_perl. The implementation architecture builds on a relational database (Postgresql), an object-relational layer (DBIx::Class), a layer of "business objects" (custom ZBW::Resource::* classes, reflecting the different RDF Classes and their properties, and using RDF::Query::Client for Linked Data access over the web), a controller component (CGI::Application, with URI mapping carried out by the CGI::Application::Dispatch::Regex Plugin) and a view component, where the actual RDFa pages are produced (HTML::Template). The user interface is made up with the YUI CSS and Javascript framework.

The universal structure of nested OAI-ORE aggregations facilitates re-use of code through methods like get_children_data() which can be applied to every level of the aggregations.

Licensing

Adding licensing information to the data set is still an open issue. While the ZBW owns the rights for the metadata (mainly the structure of the collection) and could therefore put it under an CC0 license, intellectual property rights for the main contents, namely the scanned documents, are distributed over presumably tens of thousands of authors (frequently referenced only by initials or not at all in the documents) from more than a thousand sources (most of which have long since disappeared) under dozens of different legislations. These rights can’t be granted by

ZBW at all. It's an open issue how this situation can be expressed formally and unmistakeably. Therefore, currently only an informal license statement exists.[12]

Conclusion

For 20th Century Press Archives, Semantic Web and Linked Data Techniques enhanced largely the accessibility of a large and deeply nested collection of digitized documents. Our OAI-ORE/RDFa based approach integrated well with non-semantic-web tools, and pulling in data from other Linked Data sources added considerable value for the users.

In many archives (for instance literary estates or government files) metadata is even more sparse than in 20th Century Press Archives, whereas preservation of the original order is most significant. Therefore, OAI-ORE is considered an option for expressing archival finding aids also in classical archives [6]. The experiences gained with P20 may encourage such approaches.

References

1. Huck, T.S., Wannags, M.: Die Pressearchive von HWWA und ZBW - Retrodigitalisierung der Altbestände von 1900 bis 1930. In: Burckhardt, D. (ed.) .hist 2006: Geschichte im Netz: Praxis, Chancen, Visionen: Beiträge der Tagung .hist 2006, pp. 430-445, Berlin (2007).
2. Open Archives Initiative Protocol - Object Exchange and Reuse, http://www.openarchives.org/ore/.
3. Neubert, J.: Bringing the "Thesaurus for Economics" on to the Web of Linked Data. Proc. WWW Workshop on Linked Data on the Web (LDOW 2009), Madrid, Spain. (2009).
4. Habing, T., Cole, T.: Candidate approaches for describing ORE Aggregations in METS, http://ratri.grainger.uiuc.edu/oremets/.
5. Sanderson, R., Llewellyn, C., Jones, R.: Evaluation of OAI-ORE via large-scale information topology visualization. Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries. pp. 441-442ACM, Austin, TX, USA (2009).
6. Kaplan, D., Sauer, A., Wilczek, E.: Archival description in OAI-ORE. Presented at the Open Repositories 2010, Madrid (2010).

Submission to the *Semantic Web Challenge 2010* at the *9th International Semantic Web Conference (ISWC2010)*, Shanghai, China, 7-11 Nov 2010

[12] http://webopac0.hwwa.de/digiview/digi_eigenesache.html (in German)

Appendix

Minimal requirements

1. The application has to be an end-user application, i.e. an application that provides a practical value to general Web users or, if this is not the case, at least to domain experts.

The target audience of the application is are scholars and students in economic and contemporary history, archivists, journalists and the general public. The application is already integrated with ZBWs Information Desk.

2. The information sources used

- should be under diverse ownership or control

Information sources are controlled by ZBW, Wikipedia/DBpedia, German National Library, OCLC and Library of Congress.

- should be heterogeneous (syntactically, structurally, and semantically), and

Sources are available as relational database, RDF/XML online, RDF NTriples dump or SPARQL endpoint; different structures and vocabularies are used.

- should contain substantial quantities of real world data (i.e. not toy examples).

250,000 documents with 1.8 million page image files

3. The meaning of data has to play a central role.

- Meaning must be represented using Semantic Web technologies.

RDFa web application (using ore, dcterms, skos, rdaperson, exif and a custom vocab)

- Data must be manipulated/processed in interesting ways to derive useful information and

Access to every level of large and deeply nested aggregations; data enrichment by consuming linked data; extended lookup through synonyms from LOD sources; dynamic ore aggregations for search results; integration of a legacy METS/MODS viewer.

- this semantic information processing has to play a central role in achieving things that alternative technologies cannot do as well, or at all;

Minting URIs (for citations and linking), accessing the aggregations on the web with all their subordinate parts and precisely harvesting data from them would be impossible without a vocab such as OAI-ORE. Data enrichment with web sources would be much more laborious.

Additional Desirable Features

- The application provides an attractive and functional Web interface (for human users)

The application offers a straightforward web interface with searching, browsing and viewing capabilities for users who generally will not be aware of the Semantic Web background.

- The application should be scalable (in terms of the amount of data used and in terms of distributed components working together). Ideally, the application should use all data that is currently published on the Semantic Web.

Already works for the full currently available P20 dataset. The number of documents will grow to 5.7 million over the next years. For scalability, we will add standard web caching for incoming (remote LOD) and outgoing data. The application itself can be deployed easily on multiple web servers.

- Rigorous evaluations have taken place that demonstrate the benefits of semantic technologies, or validate the results obtained.

The application is quite new, however it's a Use Case for W3C Library Linked Data Incubator Group[13].

- Novelty, in applying semantic technology to a domain or task that have not been considered before

To my best knowledge, this is the first application of OAI-ORE to large archival collections, and the first time dynamic OAI-ORE aggregations are introduced for search results.

- Functionality is different from or goes beyond pure information retrieval

The main functionality is structuring and providing access to these large piles of scanned page images.

- The application has clear commercial potential and/or large existing user base

To my best knowledge, 20th Century Press Archives is unique in offering public access to the thematic dossiers of large general interest press archives.

- Contextual information is used for ratings or rankings

not yet

- Multimedia documents are used in some way

The archives main contents are page images; for Biographical Dossiers thumbnails from DBpedia are added.

- There is a use of dynamic data (e.g. workflows), perhaps in combination with static information

not yet (we think about adding annotation features in the long run)

- The results should be as accurate as possible (e.g. use a ranking of results according to context)

doesn't apply here (no reasoning)

- There is support for multiple languages and accessibility on a range of devices

RDFa pages in English and German, switchable on every page.

[13] http://www.w3.org/2005/Incubator/lld/wiki/Use_Case_Publishing_20th_Century_Press_Archives