Metadata Management, Swiss National Licences, Mid-project report November 2015 - June 2016

Lionel Walter, Project Swissbib, Basel University Library, June 22nd 2016
lionel.walter@unibas.ch

1. Introduction

In 2015, Switzerland launched a Swiss National Licences project. This is a 2-year project, funded by swissuniversities / program SUK-P2 "Scientific information : Access, processing and storage" with a total amount of 10 million Swiss Francs. 7.5 million to buy contents from publishers, 2 million to ensure preservation for the whole Switzerland (probably with Portico and LOCKSS) and 0.5 million for the negotiations of contracts, the overall management and the metadata management. The project is led by the Consortium of Swiss Academic Libraries (at ETH Zurich). The metadata management subproject has been allocated to the Swissbib team at the University of Basel. The subproject started on November 1st 2015 and the end is running until December 31st 2016.

Access to the content licensed will be possible for all partners of the Consortium : universities, universities of applied sciences, state libraries, research institutes… One of the goal is therefore to bridge the gap between "rich" and "poor" institutions in Switzerland. Interested private persons living in Switzerland can also access the content, directly from their home, after following a suitable online registration.

The goals of the metadata management subproject are the following :

- **Private users**. Build a search engine to allow private users in Switzerland to search and access the content which is licensed for them. This will be done using the Swissbib existing infrastructure. The registration and authentication mechanism will be created together with SWITCH, as part of the Swiss edu-ID project.
- **Integration in Library Discovery tools**. For participating libraries which already have some kind of discovery tools, the integration of content should be seamless, for example with the creation of dedicated targets in ExLibris SFX Knowledge Base.
- **Gain experience in article metadata management**. The management of publisher's metadata at the article level didn't take place yet in Switzerland. The goal is to gain experience with this to deliver additional services later on (for example within the SLSP project) : integration of documents in open access repositories (using green open access conditions from national licences), text and data mining, discovery tools for smaller institutions (university of applied sciences) as well as a collaboration with international partners. All metadata associated to the project will be available for interested parties in various forms.

At the end of April 2016, the first three contracts were signed with Cambridge University Press, De Gruyter and Oxford University Press.

This summary is divided in 4 parts : the work on the

- Publisher side
- Technical side
- Collaboration side

- Customer side (libraries, private users)

At the end you can find various appendixes

- Requirements for Metadata
- Requirements for the Authentication
- Results of the Quality Control
- JATS metadata sample

---

2. Work on the Publisher side

2.1. Metadata requirements

One of the first task was to set a list of requirements regarding metadata for the publishers. You can see the results in the first appendix : “Requirements for Metadata”. Some thoughts about these requirements :

- A Creative Commons Zero Licence was required on metadata. This to ensure the maximum degree of freedom with the metadata : to be able to modify it, enrich it and redistribute it to whom might be interested (libraries, researchers, funders, other repositories, …). Some use cases are the ingestion in swissbib.ch, the transformation towards Linked Open Data or the redistribution to libraries in Switzerland.
- After thinking about a list of needed fields for the metadata, we took the option to request all fields which are available on the publisher’s platform. It is far more flexible.
- Samples of metadata are not enough to assess the quality (for example in a first step Cambridge University Press sent 2 records only). We required the whole metadata to make sure that we can work with it. A lot of problems were reported in the past regarding the bad metadata quality from publishers.

2.2. Negotiation with publishers

After long negotiations with publishers, they accepted all conditions, except the metadata licence for some of them. See the summary below for details.

2.3. Metadata Quality Control

2.3.1. Overview

In October 2015, all publishers sent samples of metadata (from 2 articles for Cambridge University Press to 128 articles from Springer). This was a good start to learn about article metadata, but far than enough to assess the quality. We requested than the whole metadata that we analyzed further.

In Appendix C, you can find the results of the quality analysis of the metadata.

In general, the results of the quality analysis were good. No major problems in publisher’s metadata, although there are a lot of minor problems, errors or inconsistencies.

The metadata analysis helped as well to have a better grasp on the coverage of the national licences contracts. For example, in the case of Springer, only half of the titles of the platform were covered by the national licence.

All publishers deliver metadata in the form of one file for each article via an FTP server.

2.3.2. Cambridge University Press

Summary of the delivery

- 51’777 zip files, for a total of 1.8 Go, in 391 folders (one per journal)
- one zip file per issue
- after extraction, one xml file per article

- 1’132’710 xml and sgm files
- delivery via FTP (on swissbib.org server)
- metadata has been delivered for more than 95% of the articles in the contract

Main issues

- the metadata for 12.8% of the articles has been delivered in SGML format. Cambridge will deliver all articles in XML format during Q3 2016 (they migrate to a new platform). There are various solutions tot tackle this problem : we can either wait, transform the metadata ourself towards XML (using osx) or use the metadata from our Canadian colleagues. Most of these files have a `.sgm` extension, but 27’301 of them have an `.xml` extension. In May, we decided to transform all SGML towards XML to be able to process it as soon as possible.
- Not all metadata from the publisher platform is in the metadata (for example the online date is missing).
- Some doi don’t resolve.

Delivery format

DTD	number of documents	proportion
JATS 0.4	149’831	13.2%
JATS 1.0	70’346	6.2%
NLM 2.2	767’070	67.7%
SGML	145’463	12.8%
Total	**1’132’710**	**100%**

2.3.3. De Gruyter

Summary of the delivery

- 33’340 zip files, for a total 8.1 Go
- one zip file per issue
- after extraction, one xml file per article
- 517’000 xml files and 1000 other files (eps, gif, …)
- delivery via FTP (on a De Gruyter server)
- the fulltext is included in xml for some journals
- some supplementary material is included as well (images, tables)
- metadata has been delivered for more than 95% of the articles in the contract, except the year 2015 which hasn’t been delivered yet

Main issues

- sometimes, the article is only a table of content, front or back matters or a list of addresses. It is not always possible to identify such cases. Here are two examples : (front matter or address list)
- some doi’s don’t resolve
- for 70% of journals, the journal title is in `<abbrev-journal-title abbrev-type=full>` instead of `journal-title`
- None of the XML is valid against the DTD. De Gruyter added some internal fields (example below). It is not convenient, but we can still work with the metadata

```
  <related-article xmlns:xlink="http://www.w3.org/1999/xlink" related-article-t
ype="pdf" xlink.href="annalen-1942-jg07.pdf" />
  <post-process status="nothing-found">2014-07-13T19:39:36.929331+02:00</post-p
rocess>
  <original type="pdf" xlink.href="annalen-1942-jg07.pdf" />
```

- For 36.6% of the files, the `DOCTYPE` declaration is missing, therefore we need to guess which DTD is used. De Gruyter told us that the value should be NLM 2.2. in this case.
- metadata is quite poor : only 15% of the articles have an abstract, 9% have keywords and 18% have the author affiliations.

Delivery format

DTD	number of documents	proportion
NLM 2.2	66'758	13.8%
NLM 3.0 & 3.0.2	185133	38.3%
JATS 1.0	54'906	11.3%
No DTD given (a priori NLM 2.2)	177'161	36.6%
Total	**483'958**	**100%**

2.3.4. Oxford University Press

Oxford delivered a first set during the period 23.2.2016-3.3.2016. After analyzing it, we noticed that Oxford delivered only metadata for the years >= 1996. We requested the whole delivery on March 21st 2016. Oxford started the 2nd delivery on April 6th 2016. But they had a lot of problems with their software for the delivery of metadata. We complained a lot, but it didn't help. They went on delivering the metadata at a very slow pace. At the time of writing, the 2nd delivery is about to finish. The numbers below refer to the first delivery.

Summary of the 1st delivery

- 29'063 tar files, for a total of 4.0 Go, in 223 folders (one per journal)
- one tar file per issue
- after extraction, one xml file per article
- 624'820 xml files
- delivery via FTP (on swissbib.org server)
- metadata has been delivered for more than 95% of the articles in the contract (for the years 1996-2015)
- a handful of image files are in the delivery

Main issues

- there was some problems with the delivery (some files were only partially delivered). The problem was that swissbib FTP server was too small and Cambridge and Oxford delivered metadata at the same time. We now have a bigger FTP server but with a more complex authentication process for providers
- For some journals, some years are missing in the delivery
- some files delivered by oxford are corrupted

- some fields are missing in the metadata. For example the information if an article is free is not always present

Delivery format

DTD	number of documents	proportion
NLM 2.0 & 2.1 & 2.2	531	0.08%
NLM 2.3	610'163	97.65%
JATS 1.1d1	14'126	2.26%
Total	**624'820**	**100%**

2.3.5. Springer

Summary of the delivery

- 38 zip files, for a total of 12.1 Go
- the zip files are not related to journals or issues
- after extraction, 2721 folders are created, one for each journal. One xml file per article
- 5'228'545 xml files for a total of 50 Go
- delivery via FTP (on springer server) with daily updates
- metadata has been delivered for more than 95% of the articles in the contract
- the delivery covers all Springer platform, not only the journals which are in the national licence contract

Main issues

- the DTD used is not a standard one, it is custom-made by Springer. However it is well documented and all xml files are following the same DTD. Springer did the conversion of the files for every DTD update. Therefore, although not standard, it is easy to work with Springer metadata which is very consistent.
- some of the zip files were compressed twice by Springer
- there are empty fields in the metadata which is rather cumbersome : `<JournalSubTitle/>`
- with 5 mio articles delivered, it was by far the biggest provider. Therefore we needed to adapt the technical workflows (move to elasticsearch for example).
- a lot of fields have unnecessary spaces at the end or at the beginning : `<JournalTitle>Applied Nanoscience </JournalTitle>`

2.4. Summary

	#journals	#articles	Years	Embargo	Next addition	CC0	Metadata delivery	Quality (on 10)
Cambridge	387	900'000	1770-2015	5 years	1.1.2022 : 2016	CC0 with special mentions	26.2.16	6
De Gruyter	345	480'000	1826-2015	2 years	1.1.2019 : 2016	CC0	18.12.2015	7
Oxford	223	621'000 (only >=1996)	1895-2015	3 years	1.1.2020 : 2016 (to be confirmed)	CC-BY-NC	1.3.2016 (only >=1996, still waiting for the rest)	6
Springer	1022	2'635'309 (out of 5'228'545)	1832-2004	10 (11) years	1.1.2017 : 2005+2006, 1.1.2018 : 2007	restrictive licence from Springer	18.1.2016	9

Fields presence (in percentage of the total number of articles)

	Cambridge	De Gruyter	Oxford	Springer
abstract	41%	15%	60%	65%
affiliations	58%	18%	83%	78%
contributor	80%	73%	92%	90%
emails	0.3%	7%	3%	41%
keywords	10%	9%	37%	45%
license-type	0.2%	5%	3%	99%

3. Work on the Technical side

3.1. Metadata

The code for this part has been released on Github as well as the technical documentation.

3.1.1. Introduction

As a first step, different tools were studied to see if they can fulfil our needs :

- Swissbib Content Collector to gather data via OAI-PMH or webdav
- OSX to transform SGML files to XML files
- OpenRefine to analyze contents and improve the data quality
- Elasticsearch to analyze the data quality
- Metafacture to transform the metadata
- PANDAS to analyze the data
- Shell scripts to unzip, rename, merge, move files and analyze the data
- SOLR to analyze the data quality
- xmllint to check xml well-formedness and validity
- XSLT to transform the metadata

3.1.2. Workflow for the analysis

After that, we decided for a specific workflow to analyze the metadata quality and export title lists

1. extract the files from the zip and tar files (via shell scripts)
2. do a first analysis with shell scripts (which DTD are used, which encoding, …)
3. create one xml file per journal (rather than one per article), it makes everything much faster afterwards
4. with metafacture, transform the metadata into JSON (either taking everything or some specific field)
5. index all data in elasticsearch
6. analyze the data in elasticsearch
7. export the results and the title lists with python

3.1.3. Pivot format

The goal is to transform all incoming metadata into a common format (pivot format). Once we have the metadata in the pivot format, we can process them using a unique workflow.

We analyzed potential candidates for a pivot format : MODS, NLM JATS, MARC21, DataCite Metadata Schema, …

We decided to use NLM JATS format. Here are the main reasons

- it's a NISO standard
- a lot of publishers already use this format (5 out of 8 in our first list, 3 out of 4 in the short list)
- it's supported and developed by the NLM (National Library of Medicine) in the United States, which is a big organization
- it has been used as a pivot format for years by the Scholarsportal from Toronto, one of the major project in this field

The main disadvantages

- It's designed for journal articles (there is a similar one for ebooks but it is not the same). MODS is for example more generic
- It's quite wordy, which could make the processing slower. It makes a heavy use of xml attributes, which make it sometimes a bit more complex to process

3.1.4. Expected workflow for the import into swissbib

We also established a workflow for the import of the metadata into swissbib.

1. **[same as above]** extract the files from the zip and tar files (via shell scripts)
2. **[same as above]** do a first analysis with shell scripts (which DTD are used, which encoding, …)
3. **[same as above]** create one xml file per journal (rather than one per article), it makes everything much faster afterwards
4. with metafacture coupled with XSLT, transform the metadata towards the pivot format NLM JATS
5. with fcv stylesheets (OCLC proprietary language), transform the metadata to pica+ format and import it in OCLC CBS document store
6. **[same as swissbib standard workflow]** export pica+ to solr format for swissbib search engine (using content2SearchDocs)
7. **[same as swissbib standard workflow]** use VuFind as the front-end : http://swissbib.ch

---

3.2. Authentication

There were two main questions on this topic :

1. How can private users authenticate themselves on publisher's platforms ?
2. How can we verify that they live in Switzerland, as required by the contracts ?

The starting point was a meeting with Switch on February 2nd 2016. This made it possible to write requirements for publishers regarding these topics.

The answer to the first question was the following : private users will use the new Swiss edu-ID Identity Provider from Switch to authenticate themselves (using SAML/Shibboleth). The publishers need to support the value `urn:mace:dir:entitlement:common-lib-terms` in the SAML attribute `eduPersonEntitlement`. Beside that, the Swiss edu-ID Identity Provider is already supported by swissbib.

Another option considered was to set-up a proxy (for example EZ Proxy) for the national licences contents. This was the option chosen by the German National Licences. The main drawbacks of this solution are :

- need to build and maintain an EZProxy server
- private users can not authenticate themselves directly on the publisher's platform. They need first to come to the proxy.

The main advantage is that there are no specific requirements or additional work to be done on the publisher's platform. For example, it can cover platforms which are not Shibboleth compatible.

The option with Swiss edu-ID seems better and is also better suited for the future (especially if the SLSP project uses Swiss edu-ID as well).

For the second question, we explored various ways :

- check that the IP address of the device is based in Switzerland (as is done by the Cochrane National Licence from the SAMW). This was rejected by publishers, mainly because it is then impossible to identify the misuses or to enforce non-commercial uses
- check that the person has a Swiss mobile phone number by sending an SMS with a code
- check that the person has an address in Switzerland by sending a letter per post (using online services like http://pinggen.com)

After negotiations with publishers, they agreed to a 14 days temporary access based on the verification of a Swiss mobile phone number and a permanent access delivered checking the residency per post. This way, immediate access is possible for a user which wants to read the content at the moment he discovered it.

We had then two additional meetings with Switch to see if the Swiss edu-ID framework fulfills all needs of the Swiss National Licences. The conclusion was that the main authentication features (user management, passwords, verified mobile phone number, verified post address, attribute `eduPersonEntitlement`) are dealt with by Switch. But some of the requirements are specific to National Licences. Indeed, the following conditions need to be met to access content from Swiss National Licences :

- User needs to accept terms and conditions for Swiss National Licences
- User must have a Swiss edu-ID account
- User must have in its Swiss edu-ID account a verified Swiss mobile phone number and must have requested his unique temporary access in the last 14 days **OR** User has a verified postal address in Switzerland
- User must not have been blocked by national licences administrators
- User must have used its Swiss edu-ID account in the last 12 months

This specific logic cannot be implemented directly in Swiss edu-ID, which remains a generic identity provider. We could implement this on the new consortium website or in swissbib. As Swiss edu-ID is already integrated in Swissbib, we chose this 2nd option. The duties of Swissbib, Swiss edu-ID and the Consortium were clarified in a document.

The implementation of these features, that we will call **Registration service for the Swiss National Licences**, won't be done by swissbib, as it is not part of the subproject metadata management. Nonetheless, it was decided that swissbib will oversee this work, which will be done by a Software development company. We then wrote a detailed set of requirements and requested an offer from a software company.

Support of the various requirements by the publishers :

	SMS verification	**support of eduPersonEntitlement**
Cambridge	yes	31.5.2016
De Gruyter	yes for 14 days	yes
Oxford	yes	Q3 2016
Springer	yes	yes

4. Work on the collaboration side

4.1. Discussion with international partners

In January 2016, we had exchanges with colleagues from Canada, France and Germany dealing with similar projects. The summary of the findings has been published on swissbib blog.

4.2. Meetings

We had various meetings with partners :

- with EPFL and UniBasel e-medien teams to see what they need to integrate national licences contents in their own tools
- with Switch for the authentication and the integration of Swiss edu-ID in swissbib
- with the Linked Swissbib team members to see what are the requirements to transform National Licences metadata towards Linked Open Data
- with the Working Group National Licences to present the status of the project
- with the IDS SFX working group to see how we might use the link resolver SFX for national licences

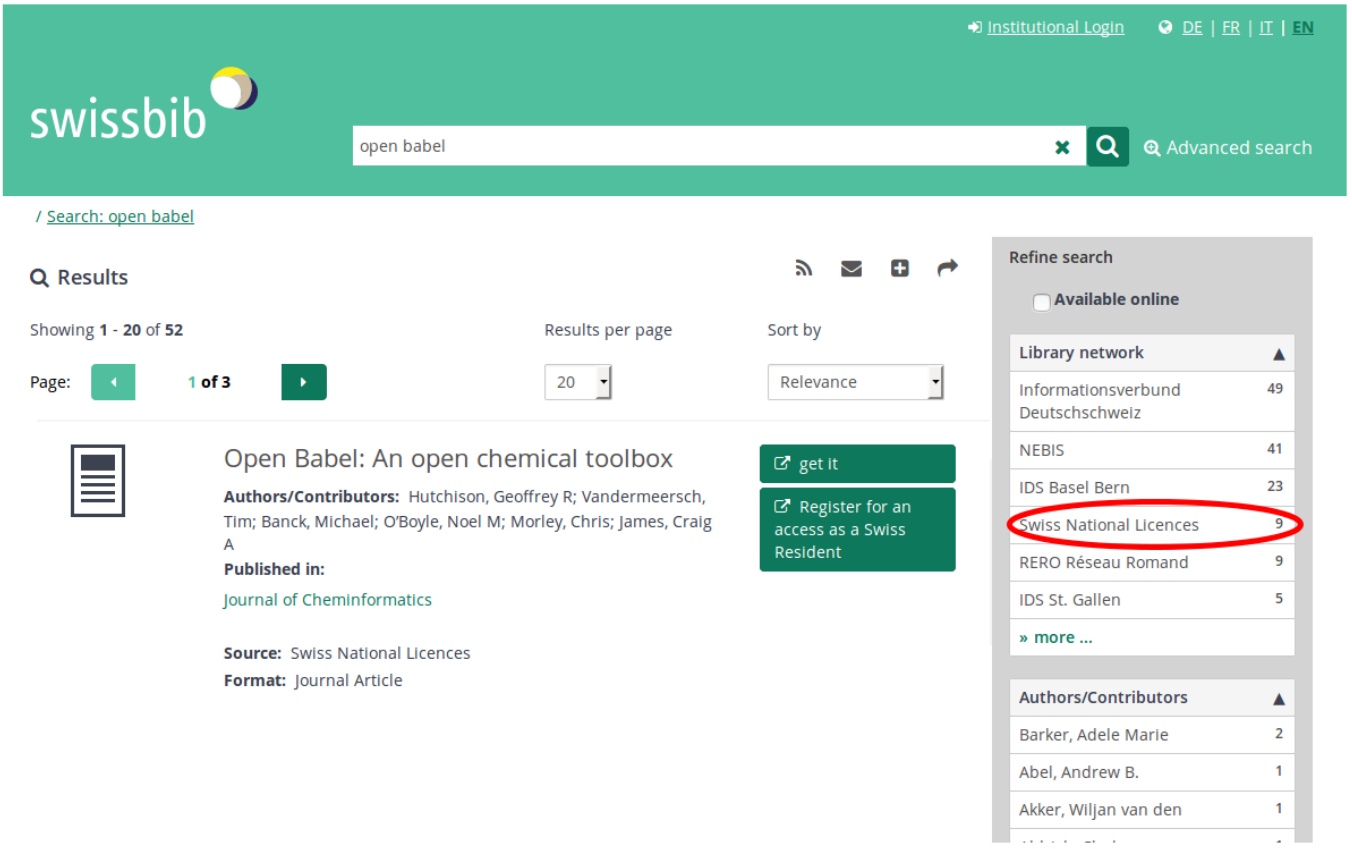
---

5. Work on the customer side

5.1. For private persons (in Swissbib)

All metadata will be integrated in the swissbib search engine. To display it to the users, various options are possible. The chosen option has a very small impact on the overall workflow, therefore it can be decided at a later time. Here are some possibilities

- integrate all articles from national licences in the standard http://www.swissbib.ch. Access to the National Licences content with a specific facet

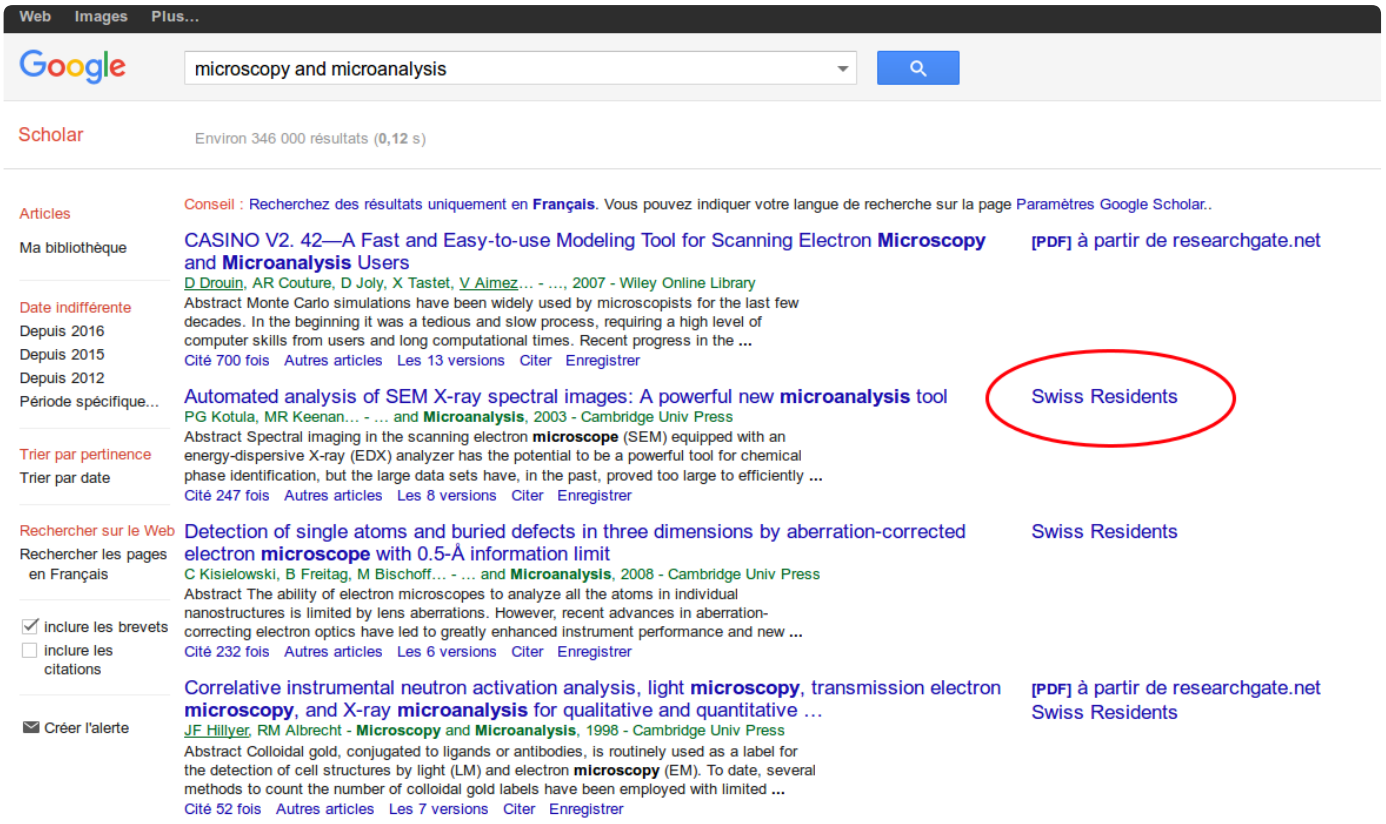


5.1.1. Other options

- integrate all articles from national licences in a specific tab from swissbib (similar to the articles tab from http://baselbern.swissbib.ch)
- take all journal articles away from the first swissbib tab and put them together with national licences in a second tab
- don't show the national licences articles in the normal swissbib, but use swissbib technology to show them on another website (for example http://nationallizenzen.ch or nl.swissbib.ch)

5.2. For private persons (in Google Scholar)

To ease the discovery of National Licences for private users and at the same time offer additional services from libraries (like document delivery), we plan to offer personalized links for Swiss residents in Google Scholar. It would look like this :



To have such a display, interested users must say in their Google Scholar personal settings that they are Swiss Residents. The most efficient way to achieve this would be to use ExLibris SFX. IDS has already a subscription to this software and it allows to achieve such processes quite easily.

A preliminary discussion took place during the last IDS SFX Working Group meeting on May 20th 2016. Participants agreed with the idea but this needs to be validated during the next KDH (Konferenz Deutschschweizer Hochschulbibliotheken) meeting on September 5th 2016.

5.3. For private persons (other means)

Some libraries in Switzerland already have some remote services for private users. For example :

- the University of Bern library has a service for people living in Bern state. Authentication technology : EZProxy.

- the ETH library has a service called E-Lending for its external patrons. Authentication technology : Shibboleth via the specific Identity Provider libraries.ch
- the EPFL has a service for its alumni. Authentication technology : Shibboleth via the EPFL Identity provider.

The various institutions want that their patrons can access the content through these means as well. For Shibboleth authentication, the requirement is to be in the Switch federation which is usually the case. For EZProxy authentication, the IP address of the EZProxy server must be declared in the national licences. This is often the case, because these servers are normally in the standard IP range of the universities.

Additionally, it must be ensured that the conditions of the contracts are met by these various services (for example residency in Switzerland). Usually this is the case as well.

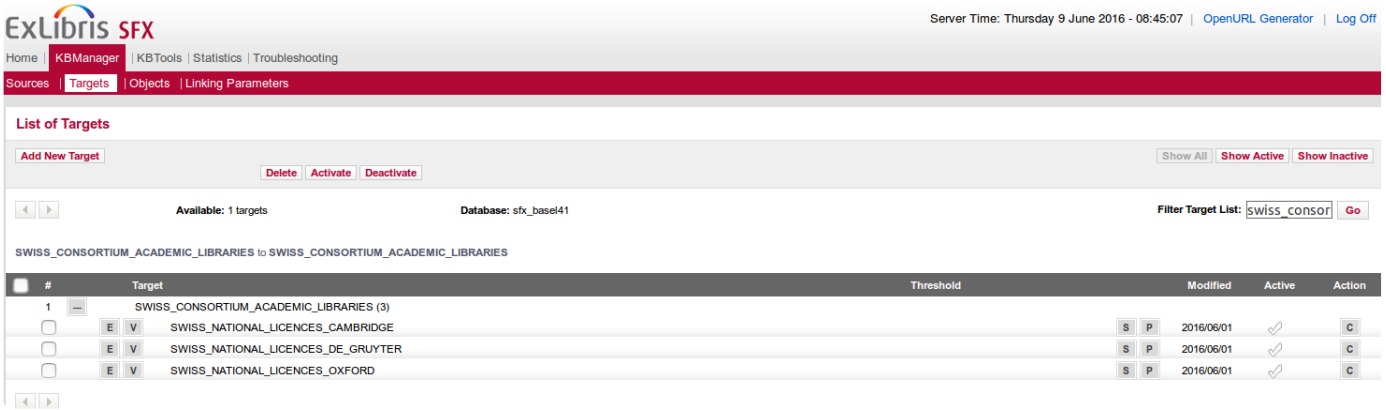
5.4. For Library Discovery tools

The first three contracts for national licences have been signed at the end of April 2016. The content was available for students and staff from all universities and similar institutions in May 2016.

Extracted from the article metadata, title lists at the journal level in Excel and KBART formats have been generated and published on the new consortium website. These lists make the integration in the various tools easier. At that point, there was quite a bit of support for the different libraries using various tools :

- ExLibris SFX and Primo (most of the Swiss universities)
- Intota (for the universities of Bern and Basel)
- EZB (University of Zurich, Zentralbibliothek Zurich, Fachhochschule Luzern, …)
- Worldcat Discovery (PH St. Gallen)
- Ovid LinkSolver (most of the universities of applied sciences)
- …

As ExLibris SFX is one of the main tool used, we took contact with ExLibris to have the possibility to create dedicated targets for Swiss National Licences in SFX KnowledgeBase. This way, the updates can be applied centrally and take effect for all libraries which use SFX. The requirement from ExLibris was to have title lists in KBART format. We complied and the targets were created at the beginning of June 2016 :



We can update Swiss National Collections centrally by depositing KBART files on ExLibris FTP server.

5.5. For Open Access repositories

The three signed contracts for national licences have green road open access clauses similar to this : “Authors from Swiss Institutions are granted permission free of charge to store their articles which are part of the Licensed Material in the form published by the publisher (e.g. PDF) in an institutional or discipline-specific repository of their choice in order to make them available in Open Access”. There is some embargo for this option :

- Cambridge : 5 years
- De Gruyter : no embargo
- Oxford : 3 years

Searching “Switzerland” in the affiliations from authors in the article metadata give the following number of articles which can be covered by this clause :

- Cambridge : ~3000 articles (published until the end of 2010)
- De Gruyter : ~1000 articles
- Oxford : ~8000 articles (until the end of 2012)

These numbers are an estimation and need to be assessed more precisely. A margin of error of 20% is estimated. After some initial discussion with the open access group of the Swiss universities, we will extract these articles and split them by institutions for the ingestion in the various open access repositories from Switzerland. The format to be delivered needs to be defined. Options might be :

- list of doi’s
- OAI set for each institutions
- other options

For the ingestion of the pdf’s, there are two options :

- download the pdf from the publisher’s platform (but there will be a watermark)
- take the pdf from the publisher’s delivery (but the publishers haven’t delivered the fulltext’s right now)

5.6. Other means of delivering the metadata

Here are what we plan to do :

- provide all article metadata with a dedicated set on swissbib OAI-PMH server
- provide article metadata via SRU and the new swissbib REST API (json compliant)
- provide all articles in pivot format via an FTP server for data mining purposes or bibliometrics
- publish it in various open data repositories (for example http://opendata.swiss)

6. Next steps and planning

6.1. Publisher's side

step	month
Import and analyze metadata from Oxford 2nd delivery	06/2016
Get 2015 content from de gruyter	06/2016
Request the missing fields with respect to the publisher's platform (cambridge)	06/2016
Analyze and report the problems with the doi's which don't resolve	07/2016
Set-up daily or weekly updates (for new content, updates or retractions)	07/2016
Set-up delivery on swissbib ftp server for all publishers	07/2016
Test completely the validity against given DTD's	07/2016
Gather the pdf's from the publishers to be able to ingest them in the institutional repositories for the "Swiss" publications	07/2016
Ensure that publisher's support authentication requirements as stated in the contract (especially the support of the SAML attribute `eduPersonEntitlement`)	09/2016
Ingest new publishers if new contracts are signed	Maybe

6.2. Technical side

Metadata :

step	month
Build the xslt stylesheets to convert all incoming metadata towards the pivot format NLM JASTS	07/2016
Build the fcv stylesheet to import in OCLC CBS	09/2016
Push all metadata to swissbib search engine	10/2016
Implement SFX linking for national licences and push subscriptions to google scholar (via IDS SFX)	10/2016
Configure swissbib search engine to handle articles from National Licences	10/2016
Update the swissbib content collector to add the support for FTP deliveries and updates	10/2016
Fine-tune the search algorithms	10/2016
Prepare the metadata and pdf's for the ingestion in open access repositories	11-12/2016
Prepare the various API's described above	11-12/2016

Authentication :

step	month
Implementation of the registration service (together with the external company) - phase 1	07/2016
Implementation of the registration service (together with the external company) - phase 2	09/2016
Implementation of wayfless links for private users	09/2016

6.3. Collaboration side

step	month
Estimate the cost and resources needed for the support of the project for 2017 onwards	12/2016

6.4. Client side

step	month
Validate (or not) the use of the guest instance from IDS SFX for the linking for national licences as well as for the displaying of content in Google Scholar	09/2016
Discuss with the open access group about the format for the delivery and the most efficient way to split the publications between the institutions	09/2016
Choose an option for the display of national licences in swissbib (tab, facet, …)	09/2016
Collaborate with the consortium for the communication	11/2016

6.5. Switch to production

If we follow this planning, national licences articles should be available for private users in swissbib on **15.11.2016**.

7. Conclusion

The work has gone smoothly up to know. The collaborations between the Consortium, Switch and Swissbib have been very good, from our point of view. The integration of the article metadata from the Swiss national licences in the various discovery tools from the universities is mostly done and this process has been simplified thanks to the KBART lists and the dedicated targets in SFX. The integration of the metadata from the articles of the Swiss national licences in swissbib is well under way and the metadata quality is satisfactory.

Here are the main risks until the end of the project :

- authentication of private users. This topic seems to be the more risky. Lot of work needs to be done in Swiss edu-ID, on the publishers' platforms and in swissbib to support the authentication of Swiss private users.
- for the open access repositories, we haven't got the pdf from the publishers yet. Therefore, there is quite a risk that these pdf will be difficult to handle.
- communication : to ensure success of the project, the communication targeted to potential private users must be good. Otherwise, the articles won't be used that much by private users.

Appendix A : Requirements for Metadata for the publishers

1. **Formats**. The Licensed Material shall be delivered to the Licensee by using open, standardized formats and accompanied by documentation.
    - Metadata for journal articles, book chapters, … : XML. A DTD or XML Schema needs to be associated with the metadata (the preferred DTD is "NLM Journal Publishing Tag Library NISO JATS Version 1.0, August 2012" or a more recent version). Other DTD or schemas are accepted if they are well documented. For each record, the associated DTD or XML Schema must be specified.
    - Metadata for ebooks and similar : MARC21/MARCXML or NLM BITS XML
    - For full texts : PDF and/or HTML/XML
    - The description of the Licensed Material (for example at journal level) shall be delivered in KBART format, with the custom coverage dates licensed.

2. **Metadata contents**. The metadata shall include all fields that are available to the publisher on its platform. This means all fields that are displayed and/or searched on the publisher's platform, including enrichments, identifiers, covers, … Later additions shall be delivered through updates.

3. **Metadata coverage**. The metadata shall be delivered for the Licensed Material in its entirety. The organization of the product into logical units (e.g. assignment of data records to products, of articles to journal titles) must be reflected by the data delivered.

4. **Delivery of metadata**. Publisher will provide the metadata through a standardized workflow for the Licensee to be able to deal with updates and deletions (for example when there are mistakes, retraction of articles, or whatever modifications could happen). This shall be at least on a monthly basis (inclusive deletions if that happens). Preferred transfer protocol is OAI-PMH, but protocols like FTP, WebDAV, or others work as well, as long as there is a way to deal with updates and deletions.

5. **Delivery of licensed material**. Publisher will provide the licensed material on a commonly agreed medium.

6. **Accessibility of licensed material**. Content has to be viewable with the usual tools like PDF-Viewer. The recommendation of the Web Accessibility Initiative (WAI) from the World Wide Web Consortium needs to be taken into account. Copy / cut / paste needs to be possible.

7. **Character set**. Metadata has to be delivered in standardized character sets (utf8).

8. **Quality control**. Before the signature of the Licence, Publisher will provide all the metadata associated to the Licensed Material for the Licensor to be able to analyze the quality of the metadata.

9. **Metadata Licence**. Publisher delivers the Metadata under a Creative Commons Zero licence (CC0 see http://creativecommons.org/publicdomain/zero/1.0/).

10. **Interoperability with software and knowledge bases from Vendors**. Publisher has to have an active partnership with Vendors of Library Technology software (like ExLibris, EBSCO, ProQuest, OCLC). Regular transfers of standardized and complete metadata shall happen in the domains of

- Link resolvers and the underlying knowledge bases : for example SFX by ExLibris or the open source project GoKB by Kuali
- ERMS tools : for example Verde or Alma by ExLibris or 360 Resource Manager by Proquest
- Discovery Index : for example Primo Central Index by ExLibris or Summon by ProQuest

11. **Documentation**. The delivery of the Licensed Material and the metadata shall be accompanied by documentation. Specifically : which DTD is used for which journals for which years ? How are the files structured (naming of the directories and the files if relevant) ? Way to deal with updates and deletions if the transfer arise via FTP ? Etc.

Appendix B : Requirements for authentication for the publishers

1. **Authentication for private persons**. The authentication for private persons will happen on a dedicated Identity Provider from the Switch federation (using SAML/Shibboleth). The users affiliated to universities will continue to use their own University Identity Provider as before. For national licences, this means that access should be granted for more than 50 IdP (up to 100 in the future). Publisher is able to deal with that. Additionally, Publisher ensures that he is able to check that the SAML attribute `eduPersonEntitlement` has the value `urn:mace:dir:entitlement:common-lib-terms` to grant access.

2. **Registration for private users**. The registration process for private users will be done by the licensor. He will check that the user has an address in Switzerland either via post or via sms on a Swiss mobile phone number.

3. **Wayfless Access to Resources**. The private users need to have the possibility to access the content with one link, without the need to select the Switch federation and the Swiss edu-ID Identity Provider via the login menu on the Publisher's platform. Therefore the publisher needs to support SP-side Wayfless url's. For more background information on this topic, see the report from the UK Federation Best Practice: WAYFless Access to Resources - Configuring on a Service and Using in a Portal

---

Appendix C : Quality Control Results

The results of the first quality analysis of the metadata have been published online :

- Cambridge University Press
- De Gruyter
- Oxford 1st delivery
- Oxford 2nd delivery
- Springer

Appendix D : sample of JATS 1.0 metadata

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE article
  PUBLIC "-//NLM//DTD JATS (Z39.96) Journal Publishing DTD v1.0 20120330//EN"
 "JATS-journalpublishing1.dtd">
<article xmlns:mml="http://www.w3.org/1998/Math/MathML" xmlns:xlink="http://ww
w.w3.org/1999/xlink" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" arti
cle-type="brief-report" dtd-version="1.0" xml:lang="en">
   <front>
      <journal-meta>
         <journal-id journal-id-type="publisher-id">LIC</journal-id>
         <journal-title-group>
            <journal-title xml:lang="en">The Lichenologist</journal-title>
            <abbrev-journal-title abbrev-type="publisher">The Lichenologist</ab
brev-journal-title>
         </journal-title-group>
         <issn pub-type="ppub">0024-2829</issn>
         <issn pub-type="epub">1096-1135</issn>
         <publisher>
            <publisher-name>Cambridge University Press</publisher-name>
            <publisher-loc>Cambridge, UK</publisher-loc>
         </publisher>
      </journal-meta>
      <article-meta>
         <article-id pub-id-type="doi">10.1017/S0024282914000401</article-id>
         <article-id pub-id-type="pii">S0024282914000401</article-id>
         <article-id pub-id-type="publisher-id">00040</article-id>
         <article-categories>
            <subj-group subj-group-type="heading">
               <subject>Short Communications</subject>
            </subj-group>
         </article-categories>
         <title-group>
            <article-title>Typification of <italic>Thelephora pavonia</italic>
Sw. and reinstatement of <italic>Cora ciferrii</italic> (Tomas.) comb. nov.</ar
ticle-title>
            <alt-title alt-title-type="left-running">THE LICHENOLOGIST</alt-tit
le>
            <alt-title alt-title-type="right-running">Short Communication</alt-
title>
         </title-group>
         <contrib-group content-type="authors">
            <contrib>
               <name name-style="western">
                  <surname>LÜCKING</surname>
                  <given-names>Robert</given-names>
               </name>
               <xref ref-type="aff" rid="aff1"/>
            </contrib>
            <contrib>
               <name name-style="western">
                  <surname>GRALL</surname>
                  <given-names>Aurelie</given-names>
               </name>
               <xref ref-type="aff" rid="aff2"/>
            </contrib>
            <contrib>
               <name name-style="western">
                  <surname>THÜS</surname>
                  <given-names>Holger</given-names>
               </name>
```

```
               <xref ref-type="aff" rid="aff2"/>
            </contrib>
         </contrib-group>
         <aff id="aff1">
            <addr-line>Science &amp;amp; Education, The Field Museum, 1400 South La
ke Shore Drive, Chicago, Illinois 60605-2496</addr-line>, <country>USA</country
>. Email: <email xlink:type="simple">rlucking@fieldmuseum.org</email>
         </aff>
         <aff id="aff2">
            <addr-line>Life Sciences Department, The Natural History Museum, Cr
omwell Road, London SW7 5BD</addr-line>, <country>UK</country>
         </aff>
         <pub-date pub-type="epub"><day>23</day><month>10</month><year>2014</ye
ar></pub-date><pub-date pub-type="ppub">
            <month>11</month>
            <year>2014</year>
         </pub-date>
         <volume>46</volume>
         <issue seq="10">6</issue>
         <fpage>825</fpage>
         <lpage>828</lpage>
         <permissions>
            <copyright-statement>Copyright © British Lichen Society 2014 </copy
right-statement>
            <copyright-year>2014</copyright-year>
            <copyright-holder>British Lichen Society</copyright-holder>
         </permissions>
         <counts>
            <page-count count="4"/>
         </counts>
         <custom-meta-group>
            <custom-meta>
               <meta-name>pdf</meta-name>
               <meta-value>S0024282914000401a.pdf</meta-value>
            </custom-meta>
         </custom-meta-group>
      </article-meta>
   </front>
</article>
```