

BHL-Europe GRIB Day, 23 February 2010, Berlin

Boris, Henning, Adrian, Heimo, Francisco, Andreas, Michael, Wolfgang, Lee, Sami, Graham, Chris

09:30

- Presentation Henning (see slides)
- GRIB is first and foremost a concept. BHL-Europes WP2 and EDITs WP5.3 search for possibilities of realization resulted in
 - o BHL-Europe bidlist [www.biologiezentrum.at:8088] based on BHL serials union catalogue/BHL bidlist
 - o demonstrator by GBV (Common Library Network) in Göttingen/Germany [http://kavia0.gbv.de/DB=1.83/SET=1/TTL=1/LNG=EN/]
- US request of data import into BHL bidlist from Prague not done. Chris gave data to Wolfgang to see what he can do with the BHLE bidlist.
- NHM does currently maintain existing BHL bidlist as it is, but no improvement done.
- Henning: GBV can offer different products.
- Henning: We have one more year with no deliverables. Boris disagrees silently, there are at least EDIT deliverables to fulfill.
- Henning: Think about alternative solutions. Chris: Curious to compare XC to GBV.
- Michael: Shows BHL-E bidlist
- Presentation Boris: Showing GRIB@GBV demonstrator and timeline
- Chris and Graham: data is the important thing, the systems around may change, the data stays.
- Transfer BHL bidlist into GBV system would replace BHL bidlist as working system
- Extensible catalogue is worth looking in, but more on a larger timescale.
- Only one realistic and pragmatic option on the table: GBV.
- However, needs lots of investigation, e.g. multilingual? different flavors of MARC? partners that have no library catalogues (xls spreadsheet)? (see *List of questions and requirements for the GRIB@GBV* at the end of this document)
- Graham: Stay with GBV at least for the next 12 months. Boris: That would be the product EDIT wants.
- Graham: GBV can remain the BHLE GRIB also for the future, and we can offer this to the global community
- How does Andreas feel seeing lots of other vendors. Andreas: trust them, impressed by stability, performance – suggests to go with them, as we have not to reinvent the wheel.
- Heimo: Java mapping tool of Vienna to use for catalogue mapping for GRIB possible?
- We can think of different products: we host the index, GBV host index, GBV build the data and works as service provider
- Graham: we don't have time to learn on the job, buy in experience from GBV helps us
- Xls sheet from Chris to Vienna to see how easy it is to get the data in the life bidlist, later shutdown NHM list?, test GBV with these data
- Integration with Goobi would save work – needs investigation
- Michael: longterm strategies, what if something happened to GBV, exit strategies
- Boris, related to what Graham said the day before: dynamic project, others (Australia, Brazil) are moving faster, our focus are the data / the content, we don't care about the platform
- Chris: Once we finished with scanning, we don't need duplication tool, we need other toolkits.

- Graham: now we have to manage a major process transforming physical objects to digital objects –looking further ahead than 2 years is not possible – we can't be strategic too far ahead because of the dynamic process.
- Talk to the librarians / scanning teams and to the technicians what they need – review. Boris: has to wait until new WP2 lead is in place.
- Lee/Sami: One time bulk data downloading not via OAI-PMH, is that possible? Work with Lucene / SOLR? Technical sheet of GBV for review by the technical team – what information do we need? How do they store the data, how we can connect to them? Good understanding of the institution GBV necessary for Atos (more political than technical question). Catalogue of requirements first and conf call afterwards.
- We have to talk to them about the connection with the different OAI modules
- Bandwidth requirements – communication with the NHM hardware system. Adrian: no problem for NHM, but has GBV enough bandwidth?
- To have data mgt. in a different side is a problem – GRIB is not data mgt.
- Where does the search happens, at the portal or at GBV, on our own system or a hosted system?
- We need to have a good degree of control of the system, otherwise it is a risk to go with GBV.
- Service level agreements – can we do the search via them....
- Graham: Contract examples available – NHM procurement team (Jim)
- If a book that needs to be scanned but is not in GRIB, it needs to be entered in the local catalogue first before a bid can be placed (*that goes back to the updating interval, how often they can update GRIB*)
- Ingest: it is stored in archival storage, please update the index in data mgt.
- Data mgt. manage OCR, name finding tools and other pieces → GBV is not going to provide this.
- Complex searching from the portal, GRIB searching for title – portal is more than the catalogue, enhanced by scan related metadata and semantical enrichment (Graham)
- Migrate indexing tools into GBV – no way to do that
- Generalising: data mgt. is the core as the data are our core.
- Merging for the portal: not yet
- GRIB is the man in the middle man between all BHL flavours on a global scale – all Scan project related searches hit GRIB first before the ingest
- Which subset of metadata goes from data mgt. to GRIB – only URL

13:30

- Identifiers: Version control is making new information available. Graham: It is related to GRIB, but more WP3, so NHM should take over it.
- Chris: CiteBank needs a GRIB, to have control. How does CiteBank feeds into GRIB?
- Tropicos: Title – reference1, reference2 (openurlresolver), CiteBank does not have that.
- GRIB has to be global, GIRBid
- BPH Number and TL2 identifiersystem for botanical literature (same as an oclc number): authority control -> Dataquality!
- Chris: Tropical datamodel works, we should use that. Chris will check his schema with the one Sami uses (might be from Kai) and sends results to Boris.
- Can BGV build the datamodel ?
- What ID to use? Doi is raped up in the publishing world, but Rod Page is in favour of that.
- How to manage unique identifier, should be between the data mgt. and GRIB
- Page level DOI?

- Open source products to be evaluated
- Handle – insert into DOI (Chris)
- How is GBV dealing with unique identifiers
- GRIB cannot become CiteBank because of the political/law background and the concept of CiteBank (uploading copyrighted content under fair use difficult in Europe)
- Resources drive the decision, we have the resources to drive GRIB, Chris has resources to drive CiteBank, we don't think about bringing it onto the same platform in the near future.
- CiteBank flat structure that manages references like Endnote or Zotero
- CiteBank has no title level authority control, that's why GRIB is required
- GRIB data needs to be reusable
- Several title ID for every title (OCLC, TL2, BHL) – can GBV provide this kind of structure – share these data (OAI?) with CiteBank
- First bulk of merging can be done automatically, manual interaction in a second step
- We have to take multiple sources input and put them into a single consolidated list that can be reused by other systems (CiteBank, Tropicos)
- Tropicos adopted BHL ID to point to the page images in BHL
- The moment we bid for a record is has to be made sure, that it is a clean master record
- We need the option to have an interface to allow manual interaction for data enhancement and cleaning records
- Graham: We need to get something done, quick, the librarians need something now, we don't have the technology experience to build this out of existing open source products, we can park this for later and/or others.
- Boris: write 1 page project plan, A4, high level, Gantt Chart with timeline and resources
- Extract original test data to compare with different systems
- NHM is minting ID and GBV needs to be able to take them in
- GRIB has to ask BHL for an ID
- Adrian will redraw the schematic overview drawn at the white board
- Library catalogues supply all data of the catalogue to GRIB
- BHLE gets the (enhanced and enriched) metadata of the scanned item with the SIP
- No blended records in BHL
- Do a list of fields that needs to go in GRIB and compare with those that needs to go in the portal – option for synchronisation between GRIB and BHLE

Summary:

- **Decision was taken to go with GRIB@GBV**
- **GRIB will stay separate from CiteBank and the BHLE OAI system, but share data with both systems – how is part of the further evaluation**
- **GRIB@GBV will be our quick working system for the libraries and the community as well as a system to fulfill the needs of the ViTaL activity within EDIT.**
- **Evaluate NHMW doing test mapping using its tool.**
- **D2.2 is finished and approved by the group, only a few minor comments to be considered by Chris and Suzanne.**
- **Further investigate unique identifiers / ID minting – WP3 activity**

Action items for Boris:

- **Send requirements / questions to GBV for feedback**
- **Send feedback to the group and set up conf call between GBV and Boris, Adrian, Lee, Chris, who else? Graham? discuss after we have received feedback**

- **Take feedback on questions and conf call and reuse for D2.3**
- **Take narrative workplan from D2.2 and add a Gantt chart with resources and send it to Graham and Chris.**
- **Provide Graham with a simple overview of the GRIB demonstrator to show the librarians.**

List of questions and requirements for the GRIB@GBV

1. Does support several languages?
 - on portal level
 - on metadata level
2. What metadata formats the GBV can handle as in- and output (MARC, etc.)
3. What file formats the GBV can handle as in- and output (spreadsheets, etc.)
4. Work with non-latin data (e.g. unicode/utf-8)
5. Library Catalogue data incorporation as well from Brazil and Australia?
6. What search engine do they use?
7. How does the database is managed / administration of the system?
8. How is GBV storing data? Type of database? Relational database? Schema?
 - Does provide technical sheets?
9. Where in the system does proprietary begin and end (only on Pica+ level?)
10. What possible products can we get?
 - Hosted solution
 - Building index of BHLE servers?
11. Is GBV working with Lucene / SOLR, can they implement this for GRIB or export data into an index?
12. What of the data / architecture are we able to take out of the GBV system later on?
13. How much of control do we have over the system,
14. How much do we have to pay and for what, how long?
15. Templates for contract from GBV or Graham
16. Sla – service level agreement
17. How often GBV is able to rebuild GRIB to update information (hourly, daily, weekly, monthly)?
18. How easily can GRIB deal with article level information
19. How does grouping of titles work / FRBRizing
20. What metadata fields should be in GRIB?
21. What metadata fields should be in the OAIS-datamanagement?
22. Possibility of harvesting selective metadata (like datasets that without scans)
23. Integration of article level metadata?
24. Can GBV build a structure that includes several title ID for one title and then share these data for other services (CiteBank, Tropicos [<http://www.tropicos.org/>])?
25. Implementation of bidding functionality?
26. Statistics about deduplication mechanism works (relation of sensitivity vs faults)
27. What algorithms do they use?
28. How does deduplication mechanism work and how can it be improved?
29. List of fields they match / matching priority order
30. How EDIT/BHL-E staff and EDIT/BHL-E partners can edit library metadata within GRIB
 - Interface for librarians allowing to clean up master records during bidding process.
31. Interface for scientist/librarians and process of manual deduplication

32. How does the library catalogue data can be incorporated automatically (what mechanisms do they offer)?
33. How does the information about e-journal licences can be incorporated automatically (what data do they need, what mechanisms do they offer)?
34. Can GBV incorporate the BHL datamodel?
35. How can we connect/integrate our tools (e.g. metadata mapping tool)?
36. How is bulk data downloading performed, except OAI?
37. How do we synchronise BHLE with GRIB (URL, ID)? NHM is minting ID and GBV needs to be able to take them in.
38. Connection to BHLE OAI datamanagement layer by synchronisation harvesting or live search?
39. On what scale do GRIB and OAI Data Management need to synchronize data? (Question to discussed with BHLE WP2 and WP3 and BHL).
40. Providing extra field for ID?
41. Possibility of requesting an id from mint system?
42. Webservices (Open url, Search titles, volumes etc. (see edit Del. 3.4 for requirements)
43. APIs?
44. Implementation of Goobi?