Minutes of Bernstein Project Meeting Liverpool, April 4th

Present: Clare Llewellyn (chair), Rob Sanderson (minutes), Marieke van Delft, Maria Stieglecker, Jeannette Godau, Thomas Fricke, Walter Schinnerl, John Harrison (Liverpool distributed database expert)

Discussion:

1. Database Mapping

The 3 fields (de en fr) are all stored in the same database. The databases are translating only the classification groups at term level. The German databases are already mapped, now mapping with WILC is ongoing. There is also discussion of additional thesaurus terms (vessel -> pot) being mapped, if possible.

2. **Specific Blueprint Changes**

Fields in only one database:
Agreement that fields in single databases are still very useful (eg estimated date in WILC).
Single database fields will be searchable, and the user will be alerted that it's only possible to perform this search in one database.
Agreement to make semantically similar but not exactly identical fields searchable through one access point, if appropriate. (eg 'estimated date' in WILC in conjunction with 'date' in other databases).
Also the opposite way – the user will be alerted if the field is not searchable in other databases, but there might be watermarks that match their query.
Discussion about the use of two pages – one where only fields available in all databases can be used, one (called 'advanced') where the user can search on any field.

Unique identifiers:
Agreement to use the name of database plus the unique identifier within the original database as the digital object identifier format, eg: info:watermarks/piccard/12345. Discussion as to whether or not it was important to use 'bernstein' as part of identifier, decision delayed until Vienna.
Discussion about the need to be able to map between identifier and human-readable citation formats.
Comment was made that it was important to use this approach as the identifiers will be in the databases eternally, and can resolve to databases even if the Bernstein workspace is no longer available.

Mapping of fields:

Piccard Online	***WZMA***	***WILC***
Ueberschriften	Motiv_long	Description
(Klassification table)	ID (only first 3 digits identical to Piccard Online)	Maingroup-descr (different system of classification) (comment 1)
Chainlines	ParA	Distance
Hoehe	parH	height
Bernerkungen	remarks	info
Aussteller		printer (comment 2)
Ort	source	place in Printer table
datierung_anfang	date_begin	date (one field for begin/end)
datierung_ende	date_end	date
Herkunft_Archiv	source (but not split)	nr_inst
Herkunft_Signatur	source (but not split)	(see comment 4)
Bernerkungen_piccard		(see comment 5)
Breite	parW	
	twin watermark field	twins
(in remarks, but not extractable)	Briquet number (generated)	nr_briquet
(resolve from Piccard number)	path to image	image (a code, not a URL)
		paper side
		paper size
		paper run
	(has similar concept, but hard to see how they could be linked)	equivalency groups
		nr_piccard (comment 6)

1. Should investigate the possibility of automatically mapping term to classification number.
2. Entity responsible for physical object (eg scribe/printer, not author)
3. Put all dates together (deduced and known)
4. Meaning: Shelfmark in archive. Discussed as possible extension for WILC, as they don't have this field currently.
5. Piccard's remarks from the physical cards. Should be merged in search with other remarks.
6. Printed Piccard numbers, so not linkable to Piccard Online. Discussion of databases working on some sort of mapping.

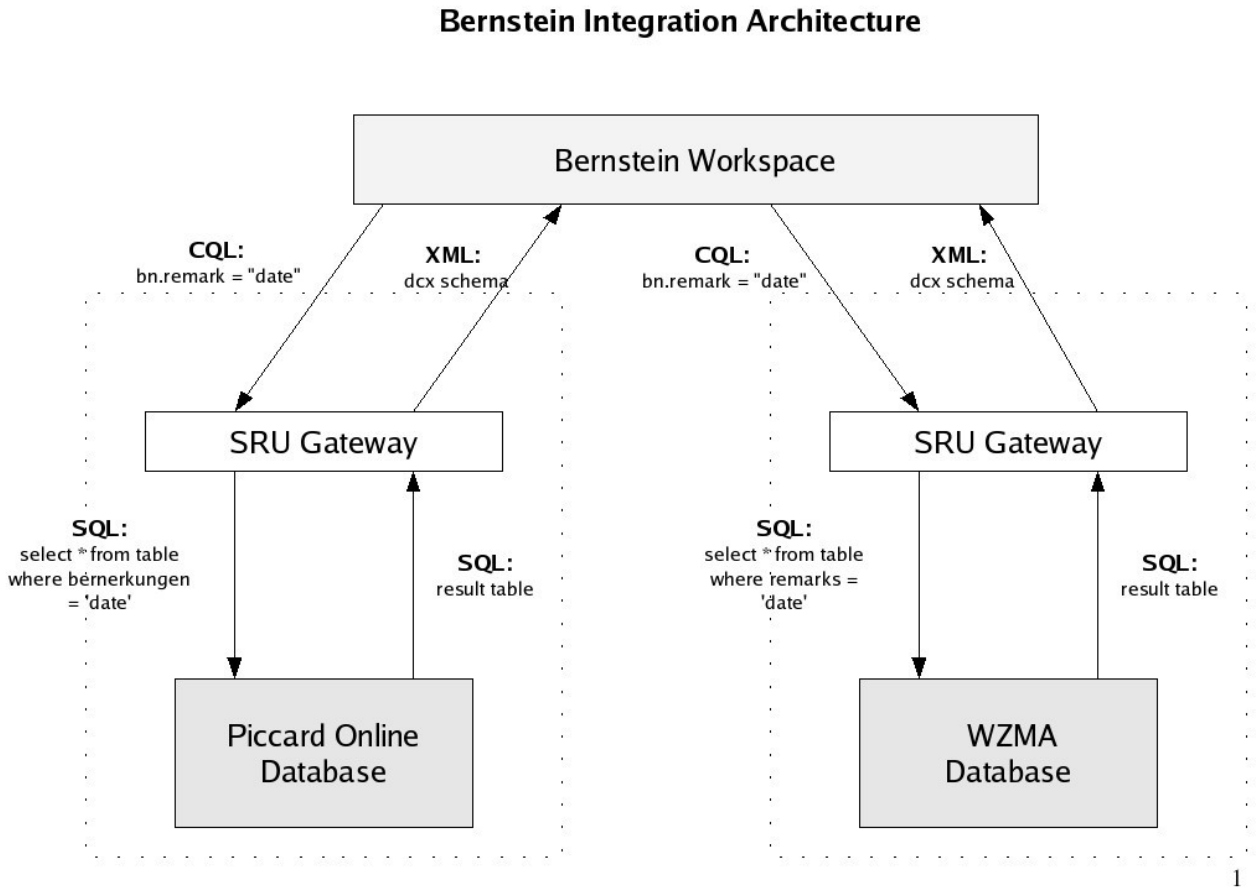
It was noted that there are 3 different types of reference field – bibliographic reference (paper source), reference to the watermark (object), and the Printed Piccard reference.

3. Linking the Databases

After extensive discussion, it was decided that all represented databases will be distributed. Unknown whether NIKI would be distributedly hosted or centrally hosted, however it was imagined that they would be centrally hosted for the sake of discussion.

Marieke described the WILC setup: Currently an access database, published/searchable on web. But the technical infrastructure at the KB is being updated at the moment to a single infrastructure. The new infrastructure will be XML based, and all records to be stored and maintained within this infrastructure. Dublin Core Extended is to be used.
Maria and Jeannette confirmed that OEAW and Piccard Online were MySQL based.

The following architecture was discussed and agreed upon:



The diagram was drawn on the board by Rob and discussed. Two databases used are examples, and can be scaled to any number of databases, including of course NIKI and WILC. The important aspects to note include that the gateway is responsible for transforming CQL into SQL, and then from the response table into XML. Each database plus SRU gateway can be hosted either by the institution or by the institution responsible for the main workspace (eg Vienna) but this location can be changed at a later stage, if required.

It was decided that SRU would be the interoperability protocol, and that the responses should therefore be in XML, in the same manner as The European Library. Dublin Core Extended was agreed upon, for similar reasons.

The following was then agreed based upon the architecture:

- Graz will provide an SRU to MySQL interface, which is configurable for CQL to various tables/columns. It will also be configurable for table/columns to XML template for the response in Dublin Core Extended (DCX).
- The databases will configure the interface for their specific setups, and host the gateway (including looking after the server machine, firewalls etc.)

Discussion then turned to the requirements for using this architecture:

- The need to agree on a CQL profile – a list of standard names of search fields. This will be discussed by the databases (plus Rob) first, and then put up on mailing list/twiki for comment.
- The need to agree upon a mapping for Dublin Core Extended XML schema. Marieke will ask Theo van Veen for mapping for DCX from WILC, and if this exists it should be used as a basis for discussion by email. (Same participants as above)
- It was decided not to discuss how the component structure will fit into the architecture as this would be more appropriate and dependent upon the meeting in Vienna in May.

For new databases to be integrated, it was discussed that the content provider should download the database and gateway software and then fill in their configuration and details. They would then contact the administrators of the workspace to ask to be linked in. The availability of SRU implementations was discussed (Bernstein specific from Graz to be produced, Cheshire3 from Liverpool, other open source implementations available)
It was considered out of scope to provide this facility for very small collections (eg 10 or so watermarks) as authentication of the information would be impossible.

4. Workplan

The following plan of work was agreed upon:

April 4th : Liverpool Meeting
April 11th: Rob to post minutes to Twiki (Rob)
April 11th: Walter to upload the latest mapping spreadsheet to the Twiki (Walter)
April 25th: Determine if date field in WILC can be split (Marieke)
April 25th: List of standard search field names (Jeannette, Marieke, Maria) (Rob to help)
May 2nd: Proposal of names to Bernstein List
May 7th: Proposal for mapping to Dublin Core Extended XML (Jeannette, Marieke, Maria)
May 10th: Vienna Meeting
July 2nd:

- Proposal for design of central workspace, including mock-up interfaces (Walter)
- Centrally hosted databases infrastructure (Walter, + Emmanuel)
- Demonstration of SRU based architecture (not integrated workspace) (Walter, RS to help)

July 9th: Fabriano Meeting
July 30th: Deliverable (as of July 2nd) to EC
September: SRU gateway version 1.0 (SRU <--> MySQL) (Walter)