BHL-E Technology Workshop Agenda
Notes (Chris Freeland)
14 May 2009

Introduction

- Motivation from EU to provide BHL-E content into EUROPEANA

Big Picture Overview

- Global architecture
  - Multiple individual repositories with redundant preservation architecture
  - BHL-E will create copy of all content digitized by partners
    - Drivers:
      - WP4 – IPR
      - Build preservation system
    - If we cannot rely on all partners to "give away" their content, it will affect architecture
    - Have to think of this as a global effort; if 2 or 3 providers in the beginning won't participate then ok
      - But, stipulated in contract that they will provide content
    - Think of the redundant repositories of backup for each institutions, for disaster recovery purposes
      - Want to avoid data loss, preserve content, protect resources
      - Recognize that each partner has a history and its own plan, but want
  - Disadvantages of proposed system
    - Some larger institutions will have to take larger role and responsibility in preservation
      - Not shared equally
      - But, allows smaller institutions to participate without overhead of building immense, expensive
    - May have to accept that some partners will only submit metadata & low resolution images
      - Incentive for participation is system architecture for backup & disaster recovery
    - Large content providers may have problems converting their information into a BHL format & repository
      - But, EUROPEANA wants persistent links and to ensure that materials will be made available
    - Need for high-bandwidth, high throughput, high processing system
      - JANET – Transatlantic bandwidth, extends into China
    - Many governments may never accept that their national content, paid for by national funds, will be distributed outside their systems

- WP4 helps here
- ***Need to look towards GBIF for model; it sits well with national governments
- Should be an opt-in model
    - Those that have the means can preserve their own content, those that don't can be assisted by the repository structure proposed

WP2
Discussion lead by Bernard Scaife
Review of Existing Serials Mashup/Bidlist

- Supplied MARC format from first 10 BHL libraries
- Used 008
- Ran crude matching scripts
    - Removed articles (the, an)
    - Tried ISSN matching
        - Few finds
    - Tried using OCLC number
        - Multiple numbers for same title in WorldCat
    - Ended up using string matching on first 60 characters in 245
- Each institution has their own login
- Algorithm identifies matches
    - Can view holdings for libraries on matched titles
    - Libraries can place a bid
        - "Bid for All"
            - That library has the entire run and will do it all
        - "Partial Bid"
            - Can bid on volumes held by library if not a complete set
    - Libraries can do manual merging/deduping
        - Find two records that appear to be the same (even with different metadata, but really are same title)
        - Take the record with the most metadata, and merge the remaining records into the new primary record
            - Any bids come along with merge

Task 2.1 – IT Development
Leader: Bernard Scaife
2.1.1 – Establish a bibliographic database system and metadata repository for monographs & serials

2.1.2 – Develop & enhance deduplication tools

- BHL has reviewed weaknesses of existing system
    - Desire to include monographs into this system
        - Existing monograph deduping system is insufficient
- Agreement that existing system is a good starting point
- How then to expand to include holdings of BHL-E partners

- Ask institutions to supply MARC records for serial data?
  - Request to manually add new records
  - Need a task group to review system & metadata provided
    - Records with titles already digitized?
    - All records from a catalogue?
    - What about existing BHL-E
    - System should be able to harvest data from other systems
      - EDIT community stores
      - OAI
      - EDIT WP5 Cybertaxonomy platform

2.1.3 – Develop a Web database to support analysis of domain content.
a) portion already digitized

b) the portion that is in the process of being digitized
- Goobi system a good candidate here
- Developed by UGOE
- Invite developers to next workshop in Leiden

c) the portion for which plans have been created for digitization
- Want a spreadsheet of titles you plan to digitize over the course of 12 months and input that into system

Task 2.2 Analysis of domain content
Leader: Tom Gilissen
2.2.1 – List of monographs & serials that are relevant for the community
- Want to encourage specialists to contribute bibliographies to help us identify & prioritize scanning

2.2.2 – Use of the Web-database to identify distribution of relevant literature

2.2.3 – Identify responsibilities for content contribution

Task 2.3 Management of the content acquisition process
Leader: Kai Stalmann
2.3.1 – Identify content holder requirements
- This is where we can address the IPR issues & requirements discussed previously about concerns from national libraries, national nodes, national repositories
- Need a discussion of what would be easiest way for you to supply content & BHL-E to ingest it
- IA harvesting process is similar
  - existing process may provide baseline for expansion
- Need for SLA (Service Level Agreement) & prioritization enforcement?
  - If all partners flood system, how you decide which gets in first?

        - Will be pure performance considerations
        - Part of WP3
    - Minimum standards are key
        - Existing
- METS key here
    - Wrapper object that encapsulates bibliographic metadata, images, structure, etc.

2.3.2 – Assist partners in implementation & evaluation of
- Scanning operations
    - Goobi
- Duplication
    - Previously described
- Discussion and distribution of data standards
    - METS
    - Citation format
    - Bibliographic metadata
    - OpenURL
- Donors & governments to facilitate the funding of scanning
- Attracting new content providers
    - Whatever is built, must assume that partners will join after start of project; systems must accommodate late entry

2.3.3 – Coordinate with EUROPEANA,
- EC metadata schema developed by EUROPEANA v.1.0 project
    - DC-based with EUROPEANA specific elements
    - Partners who have already provided information to EUROPEANA can assist here
- Need coordination with EDIT platform

2.3.4 – IPR issues
- Need to ensure that fields are present in db to accommodate work from WP4

2.3.5 – Multicultural/Multilingual aspects
- WP3 to provide guidance

Deliverables

D2.1 – Catalogue of requirements (M3, July 2009)

D2.2 – Prototpyes of dedupe tools & bibliographic systems (M9)
- Somewhat done already, but certainly needs quick work

D2.3 - Prototype of web database for content management & collection analysis
- Harvesting of files already created by EU partners

D2.4 – Content analysis & management status report 1
- Reporting phase

D2.5 Final and enhanced web-database for content management & collection analysis

D2.6 Delivery of the first version of the approved best practice guidelines & standards

- Minimum level already exists; needs quick edit, turnaround to get into partners' hands

D2.7 – Report 2
D2.8 – Report 3
D2.9 – Delivery of final best practice guidelines and standards

WP3
Discussion lead by Adrian Smales

Task 3.1 – Technological implementation
Leader: ATOS/Origin & Adrian Smales
3.1.1 – Management of the tech development team

- To be decided, depends on who are assigned to tasks
- Need to understand partner competencies, abilities

3.1.2 – Adaptation of EUROPEANA & BHL data model, workflow, harvesting, standards, specs

- Have to ensure EUROPEANA alignment
    - This is the minimum requirement, not maximum
    - Need to understand their requirements + BHL to create a superset

3.1.3 – Definition of own standards

- Data harmonization
- Each institution will have its own procedures
- Role of BHL-E is to be an aggregator for EUROPEANA
    - Have to have a large funnel & standards that get data into BHL-E that BHL-E then submits to EUROPEANA
    - Have to build lots of connectors

3.1.4 – Liase with scanning centers of national initiatives for post-processing of content

- Identify what standards are available for post-processing
- Look at ways of connecting those systems into BHL-E

3.1.5 – Development & adaptation of specific tools: name finding, OCR

- Identify & integrate emerging tools
    - IMPACT
    - Goobi
    - Nomina Name Finding enhancements
    - EUROPEANA Connect

Task 3.2 – Technical integration with EUROPEANA, BHL & national platforms
Leader: ATOS with support from AIT/Walter Koch
3.2.1 – Ensure standards for data management & image formats are consistent with international requirements for EUROPEANA, BHL

- Adoption of EUROPEANA Semantic Elements (ESE)
    - Need to understand at a detailed level these elements

3.2.2 – Develop a distributed data model that allows countries to retain control of their data while enabling distribution to international portals

- Complementary requirements
- Need a discussion of who will be the primary – home institution or BHL?

3.2.3 – Build a prototype distributed system for German language materials, and integrate with the BHL Portal

- EC concerned about EUROPEANA integration
- Need a uniquely identifiable presence for BHL-E & distribution of content
- Discussion:
    - Building another portal separate from BHL-Global doesn't make sense
    - Demonstration of localization & integration
    - Technologies used for existing BHL are first step
    - Don't want to destroy the existing system, but have an opportunity to build BHL2
    - EC clear that if they were funding this, didn't want to rebadged existing BHL
        - Can use work as starting part
    - If our baseline BHL merges with requirements, then start there
    - Hope to reduce down to a few candidate standards in coming weeks
    - We should avoid spending 2-3 person-years on rebuilding an BHL-E branch portal that duplicates existing efforts
        - Example, lessons here from BioCASE
    - Can we agree in the end on one model with multiple 'instances' that are localizable
    - EUROPEANA's software technology will be reuasable
        - Possible that it could be part of BHL-E portal
- ***MUST develop consensus on this point

Task 3.3 – Addressing distributed access and storage – long-term sustainability
Leader: Adrian Smales

- Develop a distributed access & storage system to enable national & international storage of the scanned materials; develop the storage system for long-term sustainability
    - Model presented is a first conceptual model
    - "National & International" language is important
    - Need to ensure that the underlying standard(s) are in sync throughout partners
    - "Distributed" at this level doesn't mean local providers/country-level
        - Continental or sub-continental is our scale & scope
    - Have to have an opt-in model

        - But, acknowledge that burden of adopting a new preservation standard falls back to the individual repository
        - If opt out, will BHL-E not ingest until that repository is ready
    - Priority to user
        - Finding it is more important (to user) than how it is stored over long-term
    - If an object is contributed to EUROPEANA, then the original object goes offline, the object's metadata will be removed from EUROPEANA
    - BHL-US has a model and a goal towards a preservation system, but not an existing infrastructure
    - Should consider a system with a large number of users, wide community support
        - Fedora Commons
    - What is a preservation strategy?
        - Actual physical storage
            - Today = optical
            - Tomorrow?
        - Format
            - Even if you can store or retrieve it, can you read or use the file(s)
        - Modularity
            - A strategy for putting in place technologies that can be swapped in and out as technologies emerge
    - Want to align preservation system(s) with long-standing organizations
        - History of persistence
        - Commitment to sustainability
    - In reality, can only build for the short- to mid-term (~10 years) and look towards, plan for the long-term
    - OAIS is considered to be a mandatory standard
        - EUROPEANA is not OAIS

Task 3.4 – enabling BHL Portal access in European languages – interfaces, usability, mobility
Leader: AIT/Walter Koch with support from ATOS

3.4.1 – Work with EUROPEANA to create multiple language access to the BHL Portal

- Walter has some good ideas

3.4.2 – Create a prototype portal in German to allow access to the BHL Portal

3.4.3 – Apply model to the key European languages

- Discussion:
    - Russian language was omitted
        - NHMV has Russian language experience, capabilities
    - Transliteration in the search?
        - Yes

    - Vernacular names
        - Tools like uBio/EOL will help here with species
    - Need to be aware of issues around language when aggregating via OAI

Milestones
M3.1 – Delivery of all standards, data models, technology standards, preferred technologies required for implementation (M6)

M3.2 – Technology review based on German prototype instantiation (M12)
- Portal delivered in M18

Deliverables
D3.1 Deliver composition of Technology Management Board and initial meeting (M3)
- Mostly complete
- Need to document

D3.2 Document agreed standards, best practice and system components (M6)
- Hope to have consensus by tomorrow

D3.3 Plan for managing interoperability issues, data harmonisation and the integration of the content into BHL-Europe, EUROPEANA and the BHL (M6)

D3.4 Implement plans for all components in WP3, incl. data models, technology standards etc. (M9)
- Project plan

D3.5 Technical architecture status and progress report with particular focus on the development of the German prototype (M12)
- ****Must deliver this
    - Factors into delivery of Year 2

D3.6 Release of German prototype (M18)

D3.7 Key components documented for output of D3.5 e.g. BHL-Europe Portal, OCR demonstrators, distributed storage model, etc. (M24)

D3.8 Sustainability policy for continuation of service e.g. hosting, future development, helpdesk provision for service users/content providers etc. (M30)

D3.9 Live BHL-Europe system, with distributed storage and management and appropriate tools for the continued development of services and ingress of multilingual content (M36)
- Working end-to-end system
- Ingest, output, do whatever is needed

- “Management” means processes in place for ingest, update, reuse, instantiation