ViBRANT

Virtual Biodiversity Research

Towards A Universal Bibliography – The RefBank Approach

Guido Sautter
KIT (Germany) / Plazi (Switzerland)

SEVENTH FRAMEWORK PROGRAMME

e-infrastructure

The **Bibliography of Life**

A universal bibliography of biodiversity literature

The Goal

- Collect references to all existing biodiversity publications
- Make these references available …
- … in multiple common formats

The Data Set

- Scattered over a multitude of group specific services …
- … and thousands of personal bibliographies …
- … in a multitude of formats & granularities

The Challenges

- Collect all that bibliographic data in a unified format …
- … create a sustainable open infrastructure to host it …
- … and make it available to anyone interested

Previous Approaches

… and why they did not quite succeed

- Narrow focus on taxonomic group ➔ Too few users / too little data
- Monolithic prototypes ➔ Single point of failure
- Focus on data analysis & research ➔ Insufficient interfaces, etc.
- Data curation integrated with input ➔ Contribution extremely tedious
- Several of the above

- Free commercial services ➔ Will they remain free?

The RefBank Approach

… or OK, the let's try it a different way this time

- Open coordinator free network of independent nodes …
- … that replicate bibliography data between each other

- Anyone can set up a node and link it into the network
- Strictly pull-based update propagation

- Simple data upload in multiple formats (no registration required)
- ReCAPTCHA protectes upload form from scripted uploads

- Duplicates wanted! They are the prerequisite for auto-curation
- Facilities for manual curation integrated in search interface

- Data export in multitude of formats & styles from search interface

The RefBank Approach

… or OK, the let’s try it a different way this time

- Open coordinator free network of independent nodes
  - No central authority ➔ no de facto data owner
  - Simple REST & XML based data access & exchange protocol
  - Parsed references stored as MODS XML ➔ expressive, flexible

- That replicate bibliography data between each other
  - Sustainability through redundancy …
  - … also in terms of accessibility

The RefBank Approach

… or OK, the let’s try it a different way this time

- Anyone can set up a node and link it into the network
  - Local copy possible for anyone …
  - … including dynamic updates through replication mechanism

  - Facilitates multiple implementations (current one: Java Servlets)

- Strictly pull-based update propagation
  - Directed update forwarding rather than bi-directional replication
    ➔ Each node gets to decide which others to pull updates from
    ➔ No one can compromise live data set by providing bad data
  - Also facilitates setting up “toy” nodes for research & experiments without any risk of compromising the live data set

The RefBank Approach

… or OK, the let's try it a different way this time

- Simple data upload in multiple formats (no registration required)
  - Plain text (C&P from Word), BibTeX, RIS, EndNote, MODS
  - No registration required to lower bar for contribution …
  - … but, simply give your name (or alias) so RefBank can credit you

- ReCAPTCHA protectes upload form from scripted uploads
  - Open forms susceptible to spam bots
  - ReCAPTCHA viable registration-free defense …
  - … that helps transcribing BHL data along the way

The RefBank Approach

… or OK, the let's try it a different way this time

- Duplicates wanted! They are the prerequisite for auto-curation
  - Identifying near duplicates on data import is tedious …
    … especially as it usually requires atomized references
  - RefBank avoids only character wise duplicates at this stage …
    ... safe for some whitespace & punctuation normalization
  - Redundancy facilitates eliminating typos, etc. through comparison

- Facilities for manual curation integrated in search interface
  - Correct errors as you encounter them while searching
    ➔ Data gets verified when it's used, wasting no effort

The RefBank Approach

… or OK, the let's try it a different way this time

- Data export in multitude of formats & styles from search interface
  - BibTeX, RIS, EndNote, MODS
  - Chicago, Harvard, Pensoft

    ➔ Instantly interoperable with many text processors & style templates

  - Implemented using XSLT (MODS as input) ➔ easy to add new ones

The Bigger Picture

RefBank in the context of ViBRANT's infrastructure, the GNA, etc.

- Data sets imported statically
  - **ITIS, Hymenoptera Name Server, AntCat**
- Data sets harvested periodically (harvesters upload via REST)
  - **Scratchpads**
- Data services contributing via REST
  - **Plazi, Pensoft via Plazi**
- Data sets to come
  - **CiteBank, BioStore**
- Data sets that make RefBank ultimately fly
  - **Yours! In BibTeX, EndNote, RIS, plain text, whatever format**

Future Extensions

Vision of RefBank a year from now

- Production grade learning auto-curation
- Author identification
- Add more reference styles and formats
  - For both data upload and export
  - At your suggestion
- Add ReFinder as search portal on top
  - Integrating many other data sources …
  - … importing search results directly into RefBank
- Serve as bibliography data repository for the GNA

Thank You!   Questions?

Get involved!   Upload your bibliography data!

Live System Node URLs (stable, replicating):

- http://vbrant.ipd.kit.edu/RefBank/ (KIT, Germany)
- http://plazi.cs.umb.edu/RefBank/ (Plazi, USA)

Experimental Node URL (feature pilot):

- http://plazi2.cs.umb.edu/RefBank/ (Plazi, USA)

160,000+ reference strings thus far