From Heurist to the Data Commons

Mike Lynch
Tim White

sydney.edu.au/sydney-informatics-hub

THE UNIVERSITY OF SYDNEY

Outline

- Heurist
- Overview of approach
- Heurist integration project
- Part one: from Heurist to RO-Crate
- Part two: extensions
- Digital preservation versus discovery
- What comes next

Heurist

- Web platform for building digital humanities collections
- Greater flexibility than Omeka / Wordpress, etc.
- Doing a lot of things at once:
    - Researcher workbench for building collections
    - Database for storing them
    - CMS for providing a community / public version
- Sustainability problems: not up to modern security standards (XSS)
- Small developer base
- CMS requires custom code, is not performant

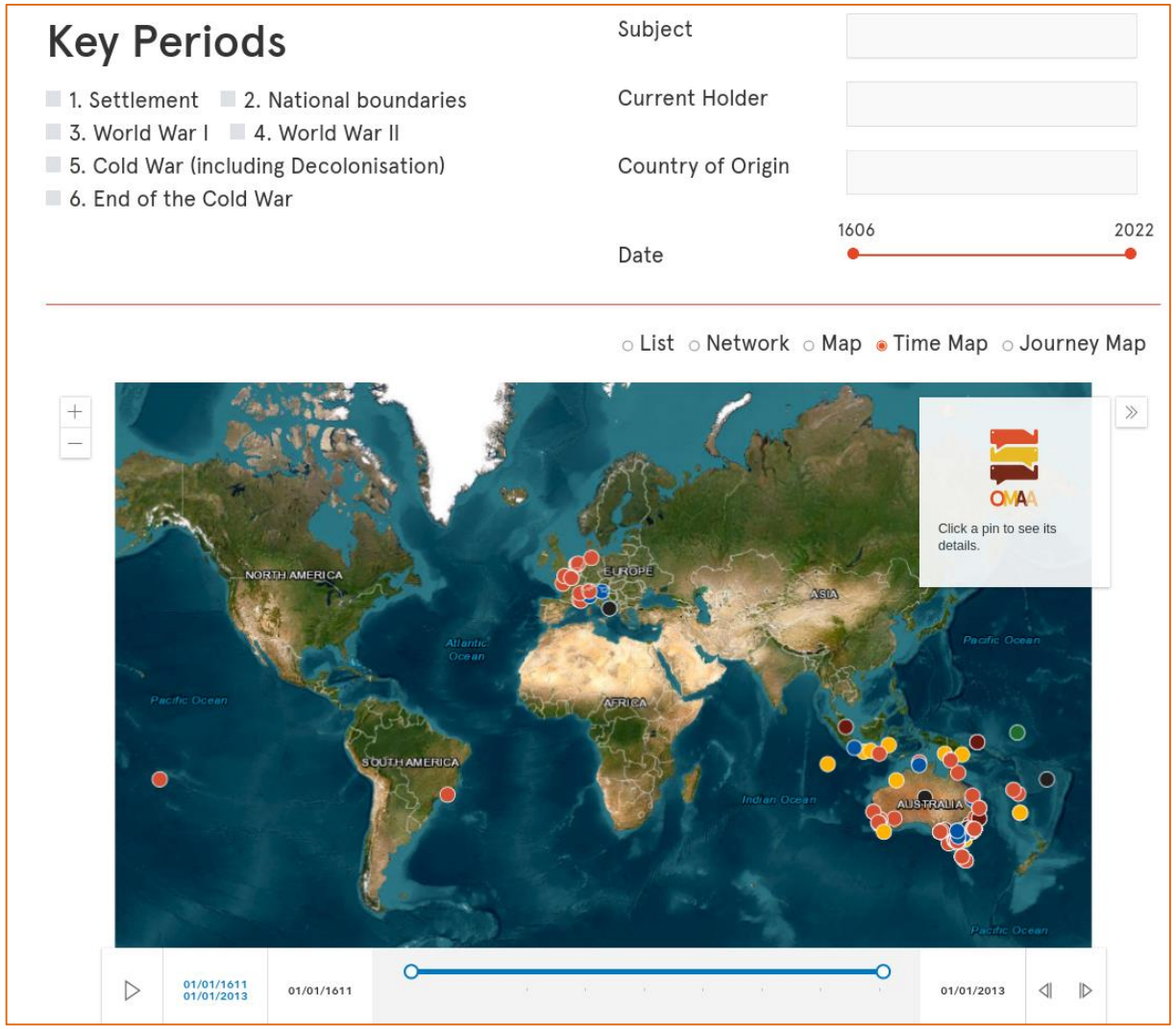
Overview of approach

- Don't build another monolith (a "Heurist replacement")
- Get collections into format which doesn't require a software stack
- Start looking at modern solutions for the use cases:
    - Researcher: "I want to build a collection"
    - Researcher: "I want to explore my collection"
    - Researcher: "I need a public-facing website"
    - Librarians, digital preservation nerds: "I want the collection to be accessible in 50+ years"
- We can't fund all of these, but breaking it down allows us to fund some

Heurist integration project

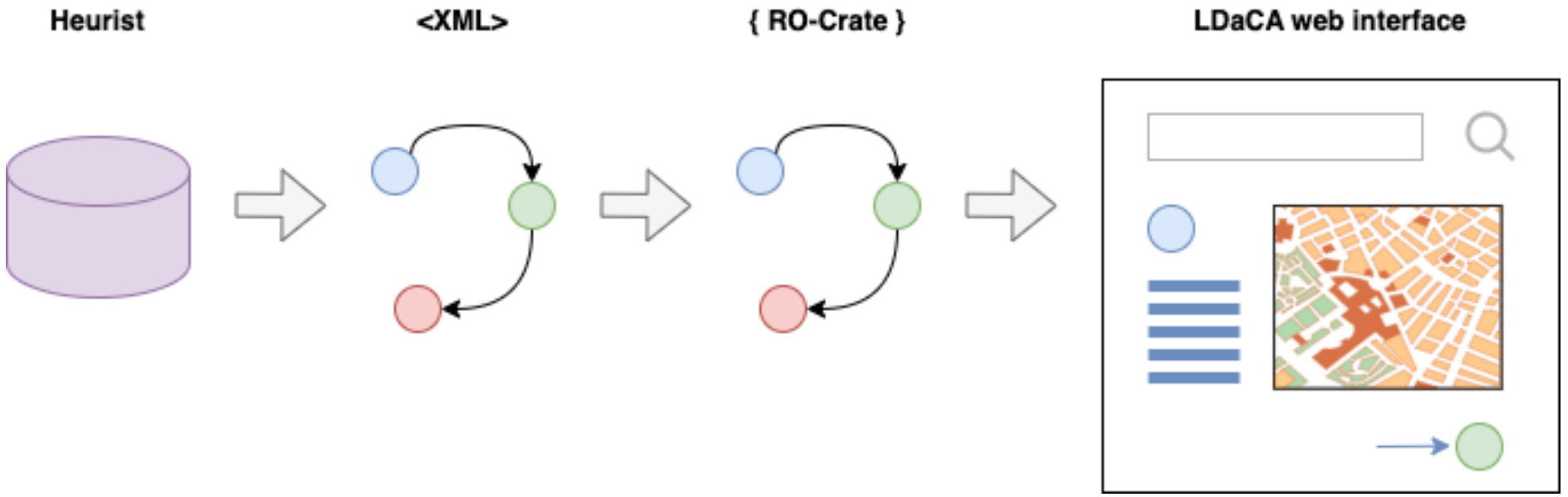
- A pilot of a broader risk mitigation strategy
- What risks?
    - Security risk posed by old software
    - Research data management risk – collections locked in old software
- Tightly scoped:
    - One major collection
    - Target RO-Crates
    - Base work on existing, ARDC-supported infrastructure

OMAA – Opening the Multilingual Archive of Australia



- A collection of historical materials in languages other than English
- Seeks to rethink and enlarge narratives about Australia
- Collection contains text-based articles, newspapers and images
- Has a Wordpress site where the archive can be explored by language, and time period: https://omaa-arts.sydney.edu.au

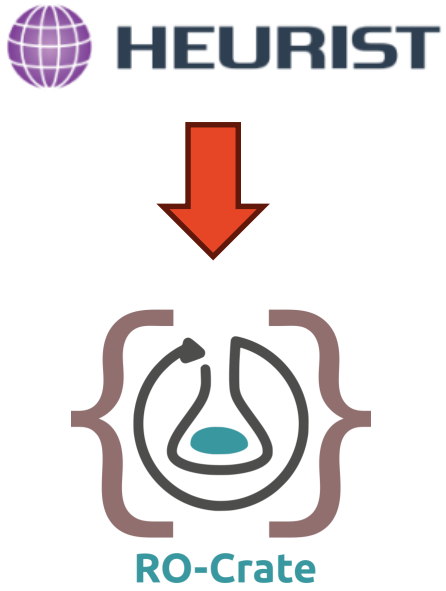
Heurist integration project



RO-Crates

- File-based standard (metastandard) for packaging research data
- JSON-LD + basic syntax
- Schema.org – basic entities like CreativeWork, etc.
- Human- and machine-readable metadata
- Easy to code against
- Existing tooling for JavaScript, Python
- In use: ARDC for LDaCA, Indigenous Data Network
- Overseas use in bioinformatics, etc.

Part one: Heurist XML to RO-Crate



- HML: Heurist's native XML export format
- Systemik Solutions – two phase approach:
  - First pass: naïve, automated mapping, get everything across
  - Second pass: have a human do some analysis, produce mappings from Heurist entities to Schema.org
- Tooling:
  - PHP for crosswalk tool
  - Crate-O and Describo – Systemik found these useful because they have schema.org classes built-in
  - Crosswalk mapping expressed as an RO-Crate

Part two: Extensions

- Oni: a discovery app for RO-Crates
  - Index one or more crates in ElasticSearch
  - A node.js / Vue SPA app allowing search and exploration
  - Currently used for Language Data Commons of Australia (LDaCA)
- What we're doing:
  - Adding maps – for a data entity, or set of search results
  - Richer geotemporal extensions using TLCMap

Search...

Collection

- [ ] Opening Australia's Multilingual Archive 1236

Member Of

- [ ] Opening Australia's Multilingual Archive 1235

Access

- [ ] Default LDaCA No License 1236

Record Type

- [ ] RepositoryObject 1235
- [ ] Text 603
- [ ] OMAANewspaper 481
- [ ] Person 40
- [ ] Place 40
- [ ] OMAAImage 30

Total: 1236 Index entries (Collections, Objects, Files and Notebooks)

Opening Australia's Multilingual Archive

Contains: Dataset RepositoryCollection

Opening the Multilingual Archive of Australia brings together historical materials from national and international collections in languages other than English. We seek to rethink and enlarge narratives about Australia that come solely from English-language sources, by showing modern Australia to be a complex multilingual creation. This project is funded by an Australian Research Council Discovery Project grant.

Language

English
Greek (Modern)
German
Dutch
French
Italian
Chinese
Japanese
Turkish
Polish
Serbian
Spanish
Arabic
Croatian
Macedonian
Indonesian
Russian
Vietnamese
Hungarian
Maltese
Korean
Czech
Danish
Persian
Ukrainian
Yiddish

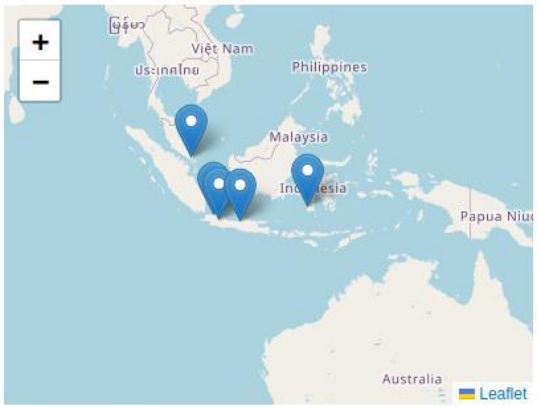
Data licenses for access

Default LDaCA No License

Objects: 1235

More

Opening Australia's Multilingual Archive > Programma der internationale voetbalwedstrijden van het vertegenwoordigend Australisch XI-tal in Nederlandsch-Indië

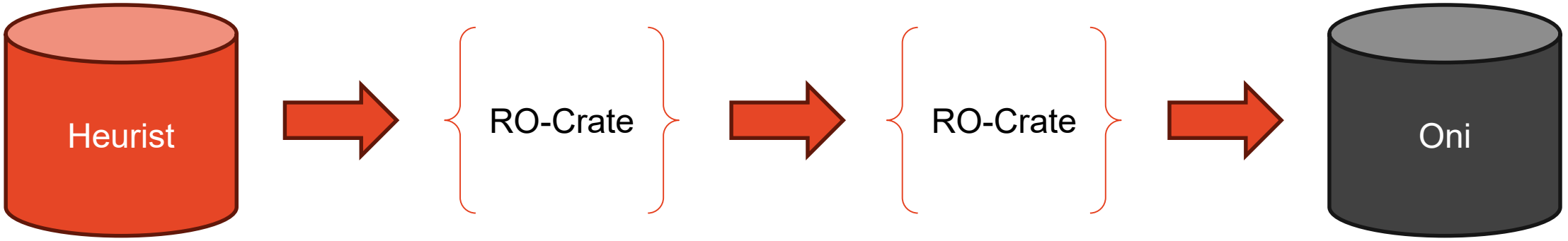
Subject	Sports
Category	2. National boundaries
Keywords	Football
Current Holder	Australian National Library
Link	https://nla.gov.au/nla.obj-107225021/view?partId=nla.obj-107225084
Bibliographic Citation	https://www.zotero.org/groups/4688363/oama/items/B2DS3N7Nhttps://www.zotero.org/groups/4688363/oama/items/MIIZ83XC
File	Football tour of NEI
Conforms To	https://purl.archive.org/language-data-commons/profile#Object
_geolocation	

Preservation versus Discovery

- Initial RO-Crate had
  - Every FOR and SEO code
  - Every Unix colour!
- Heurist XML dumps out every vocabulary it's aware of
- A preservationist approach: translate everything, because throwing things away is a risk
- This is a live collection, so a subsequent version might refer to one of those unused vocabularies

Preservation versus Discovery

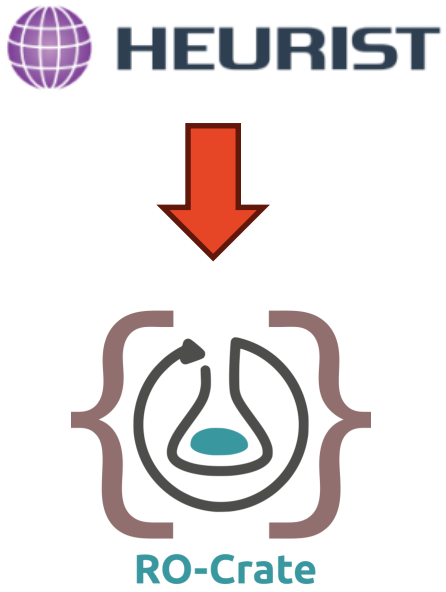
- Two use-cases for an RO-Crate
  - Absolutely everything which came from the source system (Heurist)
  - A form which can be usefully indexed in the destination system (Oni)
- Solution: decoration the preservaton RO-Crate for Oni



What comes next

- Geolocation and timeline extensions into Oni
- Working towards Oni as a full-featured discovery interface

What comes next



- Multi-stage model for the broader Heurist question
  1. Minimum: automated XML to scrappy RO-Crate for preservation
  2. Manual refinement (where funded) for better RO-Crates
  3. Decoupled platforms for
     - Public versions of websites
     - Building collections

What comes next

- RO-Crate provides bare-bones human-readable HTML
- RO-Crate with an Ansible / Terraform recipe which can
    - Spin up Oni on a server
    - Index the RO-Crate
- An archived data collection which you can deploy in minutes