# Why repository harvestability matters

UKCoRR members meeting

3rd December 2013

By Lucas Anastasiou,

KMi, The Open University

# Outline

- Exposing your metadata
- OAI organic issues
- OAI misuse
- Open Access principles
- Monitor tools
- Conclusion

# Exposing your metadata

- One of the direct objectives of maintaining a (institutional) repository is **visibility** and **dissemination**
- OAI-PMH : a lower barrier mechanism for repository **interoperability**

# Exposing your content

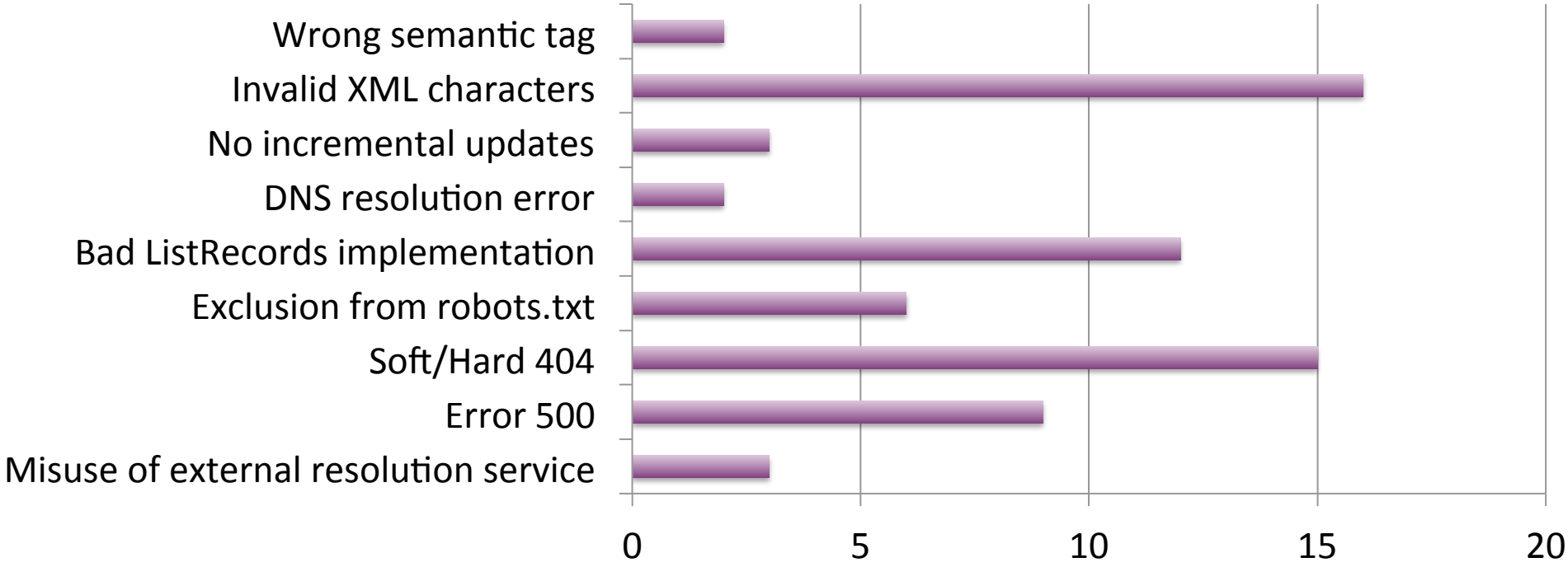
- OAI-PMH is the protocol to provide a list of your repository resources and provide incremental updates
  - Your repository's 'RSS feed'
- OAI-PMH is highly adopted
  - openDOAR reports 2508 repositories worldwide
  - Mature 'out-of-the-box' repository software supporting it: Eprints, Dspace, Bepress
- ResourceSync in the horizon
  - Do we need another standard?

# OAI-PMH organic issues

- No search mechanism : cannot filter by entity criteria
  - E.g. get me all records of author X
- Built for *metadata* harvesting not for *content harvesting*
- Weak pagination mechanism
  - resumptionTokens is a common issue
- Low granularity (per day)
- No default way to expose content in different ways according to licensing
  - though there is a workaround
- Long procedure to harvest full corpus, no way to process data on the fly, need a local copy of data to work with it

# OAI-PMH misuse

Common issues identified by CORE's 609 tracked repositories



# Open Access *content* principles

- Content referencing
  - Content referencing Open repositories should always establish a **link** from the metadata record to the item the metadata record describes using a **dereferencable identifier** pointing to the version **held in the repository**.

- Content accesibility to machine agents
  - Open repositories must provide **universal access to machines** with the same level of access as humans have.

# Common issues (1)

Exclusion from robots.txt

```
User-agent: *
Sitemap: http://repository.jisc.ac.uk/sitemap.xml

Disallow: /
```

Since OAI-PMH is used for content harvesting, exclusion from content is an (major) issue

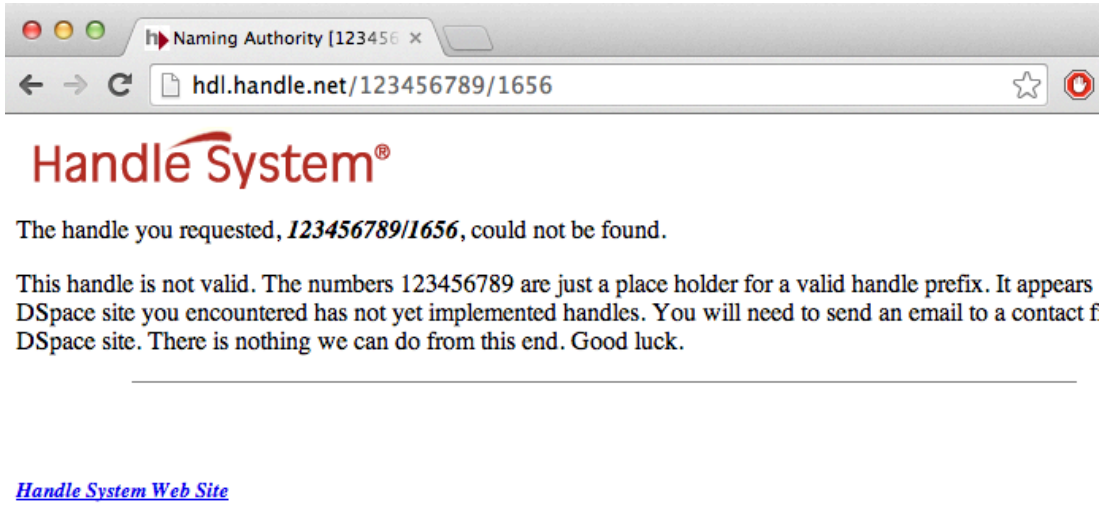
Breaks the principle of content to be universally machine accessible

Presumption of innonence.

# Common issues (2)

Misuse of external resolving service

```
. . .
<dc:identifier>
    http://hdl.handle.net/123456789/1656
</dc:identifier>
. . .
```



# Common issues (3)

## OAI Error(s)

The request could not be completed due to the following error or errors.

**Error Code** badResumptionToken

The value of the resumptionToken argument is invalid or expired.

# Discoverability / Harvestability

- Improve discoverability by service providers
- Open Access repositories registries
  - OpenDOAR
  - ROAR
- Once a document is online and Open Access it should be retrievable and can be processed by humans and machine agents as well
- Follow guidelines / best practices, avoid pitfalls

# Monitor tools

- Validation tools to check against your repository metadata/content exposure
  - CORE repository analytics
  - RIOXX Guidelines
    - http://rioxx.net/
    - Consistency of metadata fields, tracking of research outputs across scholar systems
  - OpenAIRE
    - https://www.openaire.eu/
    - OpenAIRE guidelines, OpenAIRE validator service

# CORE repository analytics

Simple information on how much content has been aggregated by CORE service, repository harvesting status and logs of harvesting attempts (with issues included)

| | | | | | |
|---|---|---|---|---|---|
| 165 | Parade@Portsmouth | ✓ (8641) | ✓ (2291) | ✓ (1956) | ★ |
| 166 | Edge Hill Research Archive | ✓ (4564) | ✓ (1502) | ✓ (185) | ★ |
| 167 | EdShare | ✓ (3232) | ✓ (3071) | ✓ (631) | ★ |
| 168 | Sunderland University Institutional Repository | ✓ (3342) | ✓ (533) | ✓ (206) | ★ |
| 169 | CURVE/open | ✓ (2740) | ✓ (2154) | ✓ (409) | ★ |
| 170 | LSE Theses Online | ✓ (664) | ✓ (666) | ✓ (666) | ★ |
| 171 | Research at the University of Wales, Newport | ✓ (411) | ✓ (409) | ✗ | |

http://core.kmi.open.ac.uk/repository_analytics

# So why does it matter after all?

- To achieve the primary goal of IR: to '*open and disseminate research outputs to a worldwide audience*'

- Provide the best quality content to **service providers** that make your repository more 'discoverable' , 'accessible', 'reusable'. Aggregators can act as ambassadors in your behalf

- Add value (**or avoid removing value**) by disseminating your content in the best possible way

# Conclusion

Currently used protocol (OAI-PMH) has limitations

- By design
- By misuse

Don't need to embrace a new standard (**yet**), make the best out of the current standard

Unleashing your data to the web, makes your organisation research output more visible, expands your audience