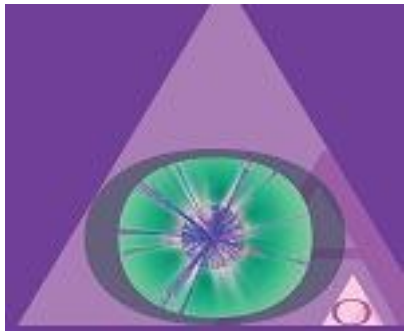

SCOAP3 Repository

Metadata handling

Agenda

1. Introduction
   - The repository view
   - Coming soon
2. Implementation details
   - Harvesting/Delivery
   - Compliance checks
   - Metadata cleaning
   - Data exposure (OAI-PMH & API)
3. Summary

1. Introduction

The repository view

Lunched at January 2014

Metadata handling process

- Ingestion/Harvesting
- Check
- Serve
- Export

Coming soon

Invenio v. 3

- Search with ElasticSearch
- JSON as main storage format
  - MARCXML still available as export and edit format
- Based on Flask microframework

2. Implementation details

Harvesting / Delivery

- Multiple ways of articles delivery
  - Periodical FTP download
  - Active push by publishers
  - OAI-PMH harvesting
- Multiple metadata XML formats (JATS, A++, Elsevier...)
- HarvestingKit - https://github.com/inspirehep/harvesting-kit
- HEPCrawl (will substitute HK)
  - https://github.com/inspirehep/hepcrawla

Compliance checks

- Contractual obligation of publishers
  - 24h delivery
  - Copyrights not transferred to publisher
  - Clear statement: "Funded by SCOAP3"

- Future use of new tools from Invenio v.3
  - https://github.com/inveniosoftware

Metadata cleaning

- Multiple automatic methods for records metadata cleaning
  - use of Bibcheck (Invenio 1.x module) to be replaced in Invenio v.3
- Outside of Invenio can be replaced by similar solutions
  - framework for scheduling and performing custom coded functions following certain interface

https://github.com/SCOAP3/scoap3/tree/master/bibcheck_plugins

Export: OAI-PMH & API

- XML based output
- In future also RecJSON

3. Summary