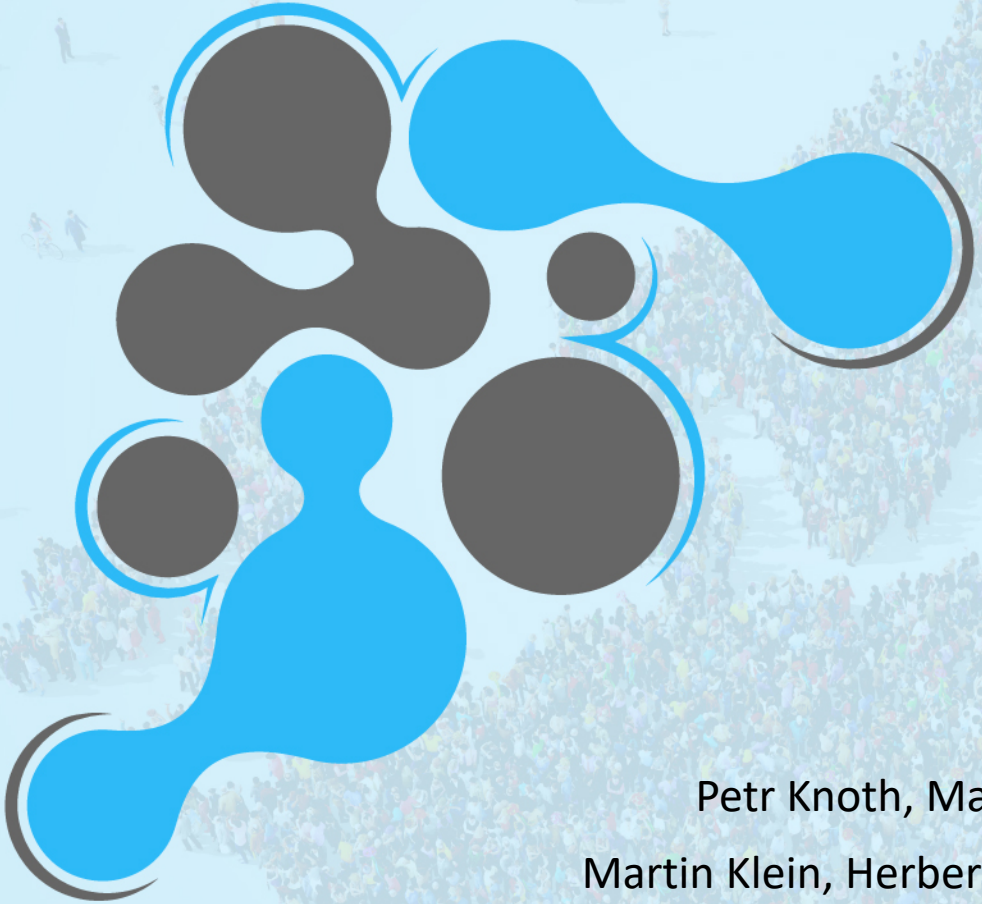

Evaluating the Performance of OAI-PMH and ResourceSync

Petr Knoth, Matteo Cancellieri,
Martin Klein, Herbert van de Sompel

The Open University, UK & Los Alamos National Laboratory, USA

EOSC pilot
The European Open Science Cloud for Research Pilot Project
www.eoscpilot.eu

EOSC Pilot

The **EOSCpilot project** supports the first phase in the development of the **European Open Science Cloud (EOSC).**

Develops a number of demonstrators functioning as high-profile pilots that integrate services and infrastructures to show interoperability and its benefits in a number of scientific domains.

The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

The problem we are addressing

- A single scientific repository is of limited value
- Real benefits often come from the ability to exchange information within a network of repositories
- Current technology for exchanging data across repositories based on a 15 year old technology (OAI-PMH).

The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

The problem we are addressing

- OAI-PMH is not
  - Not scalable for large quantitates of resources
  - Suffers from inconsistent implementations (insufficient interoperability)
  - Deals only with the transfer of metadata rather than the resources themselves

The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

Aim of the project

Assess how scientific resources can be effectively, regularly and reliably exchanged across systems using the ResourceSync protocol.

The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

Approach

Conduct a set of experiments/benchmarks comparing OAI-PMH with ResourceSync along a set of dimensions, scenarios and implementation setups.

The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

Methodology

- **Type:** baseline (batch), incremental, selective synchronisation
- **Resource type:** metadata only vs metadata and resources synchronisation
- **Implemetation:** sequential vs parallelised synchronisation
- **ResourceSync method:** single, batched and Resource Dump synchronisation
- **Performance:** speed (time), complexity (steps required to complete), reliability (recall), freshness (e.g. average achievable time gap between syncs)

The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

Activities

- Developing a scalable implementation of ResourceSync client and server
- Running experiments
- Evaluating and analysing the results
- Disseminating the results
- Reaching out to external partners to test and productionise this technology

The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

Activities

ResourceSync is a Synchronization framework for the web consisting of various capabilities that allow third-party systems to remain synchronized with a server's evolving resources.

The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

ResourceSync - Sitemap

```
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">

  <url>
    <loc>http://example.com/res1</loc>
    <lastmod>2017-01-02T13:00:00Z</lastmod>
  </url>

  <url>
    <loc>http://example.com/res2</loc>
    <lastmod>2017-01-02T14:00:00Z</lastmod>
    <changefreq>daily</changefreq>
  </url>
  ...
</urlset>
```


The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

ResourceSync – Resource List

```
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
        xmlns:rs="http://www.openarchives.org/rs/terms/">
  <rs:md capability="resourcelist"
         at="2017-01-03T09:00:00Z" />
  <url>
    <loc>http://example.com/res1</loc>
    <rs:md hash="md5:1584abdf8ebdc9802ac0c6a7402c03b6"
           type="application/pdf" />
    <rs:ln rel="describedby"
           href="http://example.com/res1_dublin_core_md.xml"
           type="application/xml" />
  </url>
  <url> ...
  </url>
</urlset>
```


The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

ResourceSync characteristics

- Explicit link between metadata and the described resource
- Designed to allow synchronisation of resources, not just metadata
- Web-centric

The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

Synchronisation approaches

- ResourceSync standard (on demand synchronisation using a ResourceList)
- ResourceSync Batch (on demand Resource Dump)
- Resource Dump (materialised Resource Dump)

The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

Differences in OAI-PMH performance

IR Software	#Repos	AVG(Rec/s)	MED(Rec/s)	σ^2
DSpace	659	147.29	71.64	1.07e+06
EPrints	402	35.48	29.14	1.98e+03
Digital Commons	149	28.57	11.47	3.70e+04
OPUS	74	39.56	23.84	2.42e+03
OJS	70	11.13	10.13	5.77e+01
dLibra	55	13.39	0.35	2.71e+03
Fedora	16	71.59	32.13	2.18e+04
Invenio	14	136.17	70.42	6.97e+04

Table 1. Comparing different IR software download speed

The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

Initial results – metadata harvesting performance

IR SW	#records	RS standard	RS batch100	RS batch500	RS batch1000	RS batch2000	RS batch 5000	OAI-PMH
Median performing repository per IR platform (based on OAI-PMH speed)								
DSpace	2751	-	141.72	315.26	616.26	707.93	702.14	71.64
Eprints	5000	-	87.28	238.23	450.86	533.81	201.60	29.12
Digital Commons	5000	-	109.18	253.92	517.54	1166.41	904.81	11.47
Opus	2500	-	184.87	134.68	799.23	815.93	805.15	23.42
OJS	1360	-	40.07	176.53	467.03	492.22	517.50	10.11
Dlibra	5000	-	70.08	490.68	1083.19	1097.16	461.89	1.47
Fedora	3997	-	63.28	253.09	744.04	741.97	432.62	29.96
Invenio	5000	-	79.09	363.50	657.89	437.54	554.02	65.59
Top performing repository per IR platform (based on OAI-PMH speed)								
DSpace	5000	-	76.95	217.16	285.71	825.54	334.72	2684.56
Eprints	5000	-	60.90	413.02	507.15	1143.73	993.64	234.22
Digital Commons	5000	-	104.57	321.52	477.10	1031.11	687.85	70.37
Opus	2220	-	166.29	773.25	242.15	691.37	735.59	228.32
OJS	1425	-	77.83	186.89	180.15	510.75	469.37	28.75
Dlibra	5000	-	84.34	357.27	611.92	453.14	367.43	176.11
Fedora	2751	-	69.83	667.07	709.75	501.92	695.22	124.67
Invenio	5000	-	88.55	228.44	204.54	508.17	409.17	120.29

Table 6. Harvesting speed of different ResourceSync implementation compared with the equivalent data harvested through OAI-PMH

The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

Initial results
requests per downloaded resource

IR Software	#Repos	σ^2(OAI-PMH)	#Requests(OAI-PMH)	#Requests (RS)
DSpace	366	8.97E+07	1529.85	1
Eprints	325	9.46E+02	8.63	1
Digital Commons	27	6.42E+08	13286.29	1
Opus	48	2.97E+05	152.33	1
OJS	49	3.52E+02	7.66	1
Dlibra	1	-	9.57	1
Fedora	15	1.50E+05	150.92	1
Invenio	6	4.46E+03	45.16	1

Table 5. Number of requests per successful document download

The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563

Conclusions

- If you are a data provider, then adopt ResourceSync
- Ongoing project, experiments to be completed by December 2018
- Outputs:
  - paper benchmarking OAI-PMH against ResourceSync across a range of scenarios
  - Scalable implementation of ResourceSync server and client

The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563