207

Publication Life Cycle at CERN Document Server

CERN Document Server hosts over 1,300,000 different documents.

How does their life looks like?

Ingestion

Data is ingested from various sources through one of the workflows:

- 100+ submission forms created during 15 years of running CDS (don't ask how we still can still maintain them) – each tailored to meet the specific needs of a given user group.
- Single- and multi-records editor – power tools for catalogers to submit the desired metadata directly.
- Automatic import from other systems, like OAI Harvesting from Inspire and arXiv, batch upload from the ATLAS experiment, etc.

Curation

Ingested and indexed data is further improved through:

- Automatic plots and captions extraction, cross-referencing with information from other collections or other systems (Inspire or ArXiV), automatic DOI minting.
- Review process, in which the internal documents, after discussions and revisions, are being publicly published.
- Authors disambiguation process that run in 2016 and, with the help of machine learning tools, enriched metadata of over 2,700,000 signatures in more that 67,000 records.
- Recommendation system that, based on how people interact with CDS, suggest relevant/ or similar documents.

Export

There are multiple ways for retrieving data from CDS:

- Each record can be exported in one of 7 formats: BibTeX, MARC, MARCXML, DC, EndNote, NLM and RefWorks.
- Each collection of records defines it's own RSS channel.
- Users can use search queries to define notifications.
- Videos can be exported to YouTube or embedded on any other website.
- With OAI-PMH protocol, other repositories can harvest records from CDS.