peer

Publishing and the Ecology of European Research

Dr Laurent Romary
Inria

PEER End of Project Results Conference, Brussels, 29 May 2012

The PEER Observatory

Supported by the EC eContent*plus* programme

PEER Observatory

- The Observatory consists of
  - Publisher platforms (usage data & access to authors)
  - PEER Depot
  - PEER Repositories
- The PEER Depot
  - Acts as a „Clearing House“ - is a Dark Archive!
  - Processes deposits and distributes content to participating repositories
- The PEER Repositories
  - Provide the usage data (= log files) needed by our research partner CIBER
- Content inflow
  - 241 journals from four broad areas (Life Sciences, Medicine, Physical Sciences, Social Sciences & Humanities)
  - 2 ways of articles deposit: publisher deposit / author self-archiving

Participating Publishers

- BMJ Publishing Group
- Cambridge University Press
- EDP Sciences
- Elsevier
- IOP Publishing
- Nature Publishing Group
- Oxford University Press
- Portland Press
- Sage Publications
- Springer
- Taylor & Francis Group
- Wiley-Blackwell

Participating repositories

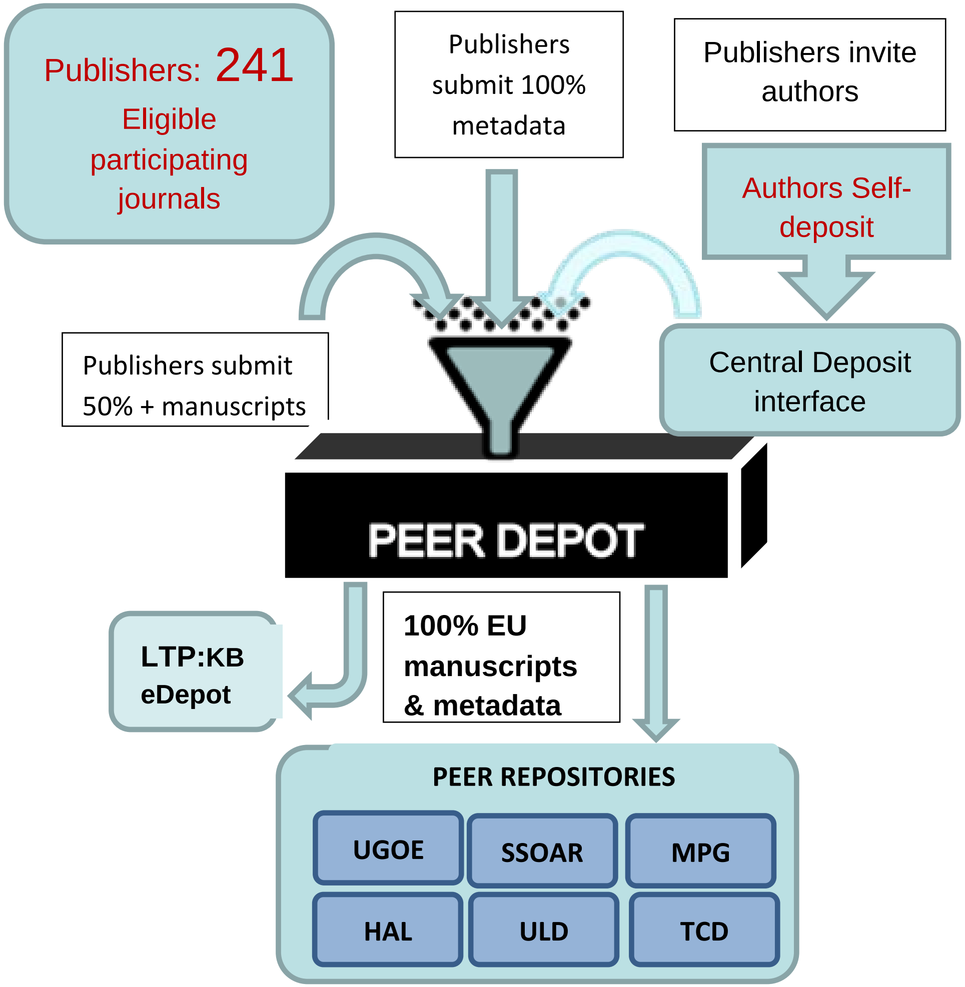
- eSciDoc.PubMan.PEER, Max Planck Digital Library (MPDL), Max-Planck-Gesellschaft zur Förderung der Wissenschaften e. V. (MPG)

  eSciDoc.PubMan

- HAL, CNRS & Institut Nationalde Recherche en Informatique et en Automatique (Inria) hal en ligne
- Göttingen State and University Library (UGOE) SUB
- SSOAR – Social Sciences Open Access repository (GESIS – Leibniz Institute for the Social Sciences) SSOAR Social Science Open Access Repository
- TARA – Trinity College Dublin (TCD) TARA Trinity's Access to Research Archive
- University Library of Debrecen (ULD) DEENK www.lib.unideb.hu
- *Long term preservation archive:* e-depot, Koninklijke Bibliotheek

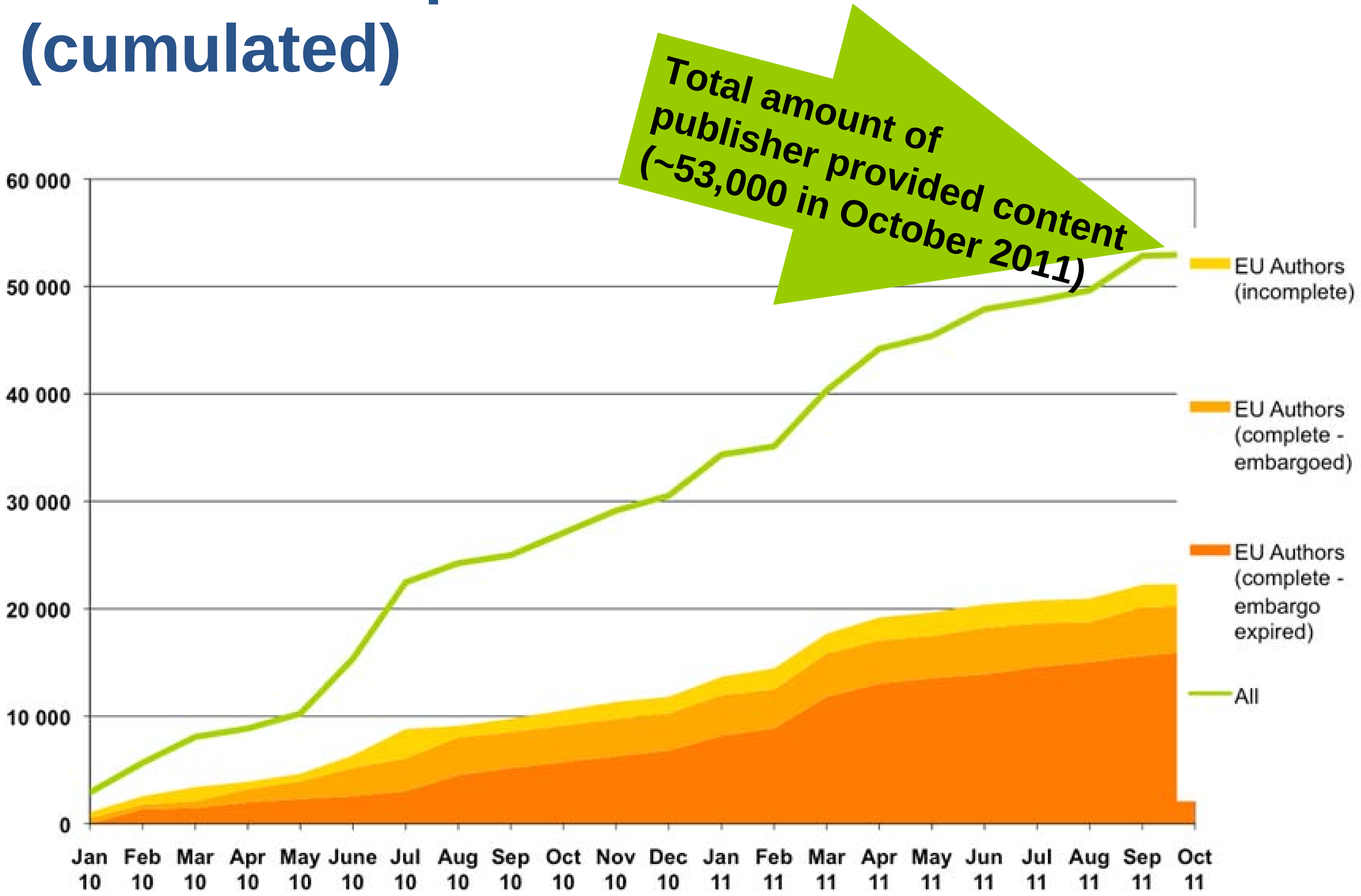
  e_Depot KB Koninklijke Bibliotheek Nationale bibliotheek van Nederland

The PEER Observatory – content flow


Publishers: 241
Eligible participating journals
Publishers submit 100% metadata
Publishers invite authors
Authors Self-deposit
Publishers submit 50% + manuscripts
Central Deposit interface
PEER DEPOT
LTP:KB eDepot
100% EU manuscripts & metadata
PEER REPOSITORIES
UGOE
SSOAR
MPG
HAL
ULD
TCD

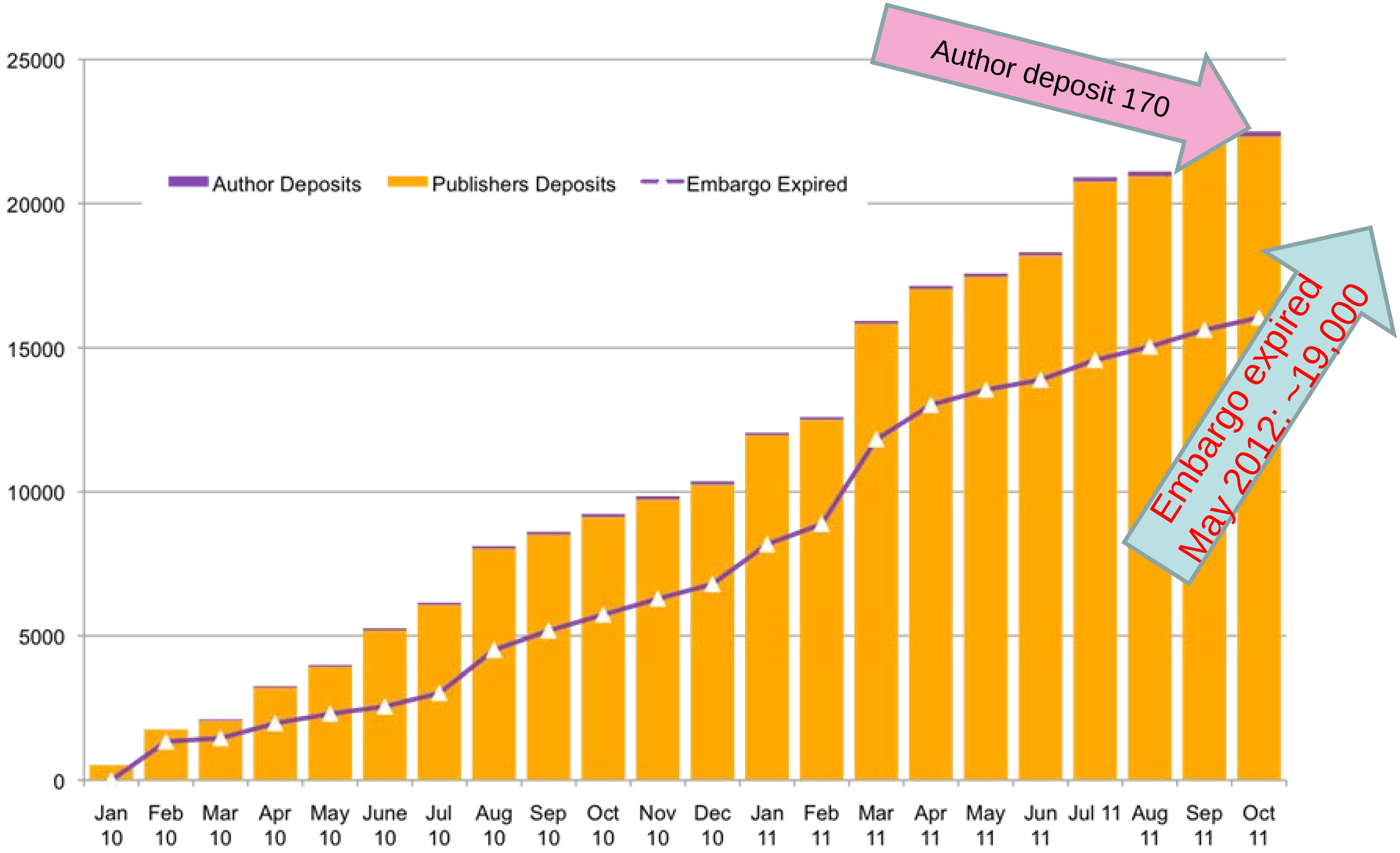
“Observatory” developed to monitor the impact of systematically depositing stage-two outputs on a large scale

Publisher deposits (cumulated)



Total amount of publisher provided content (~53,000 in October 2011)

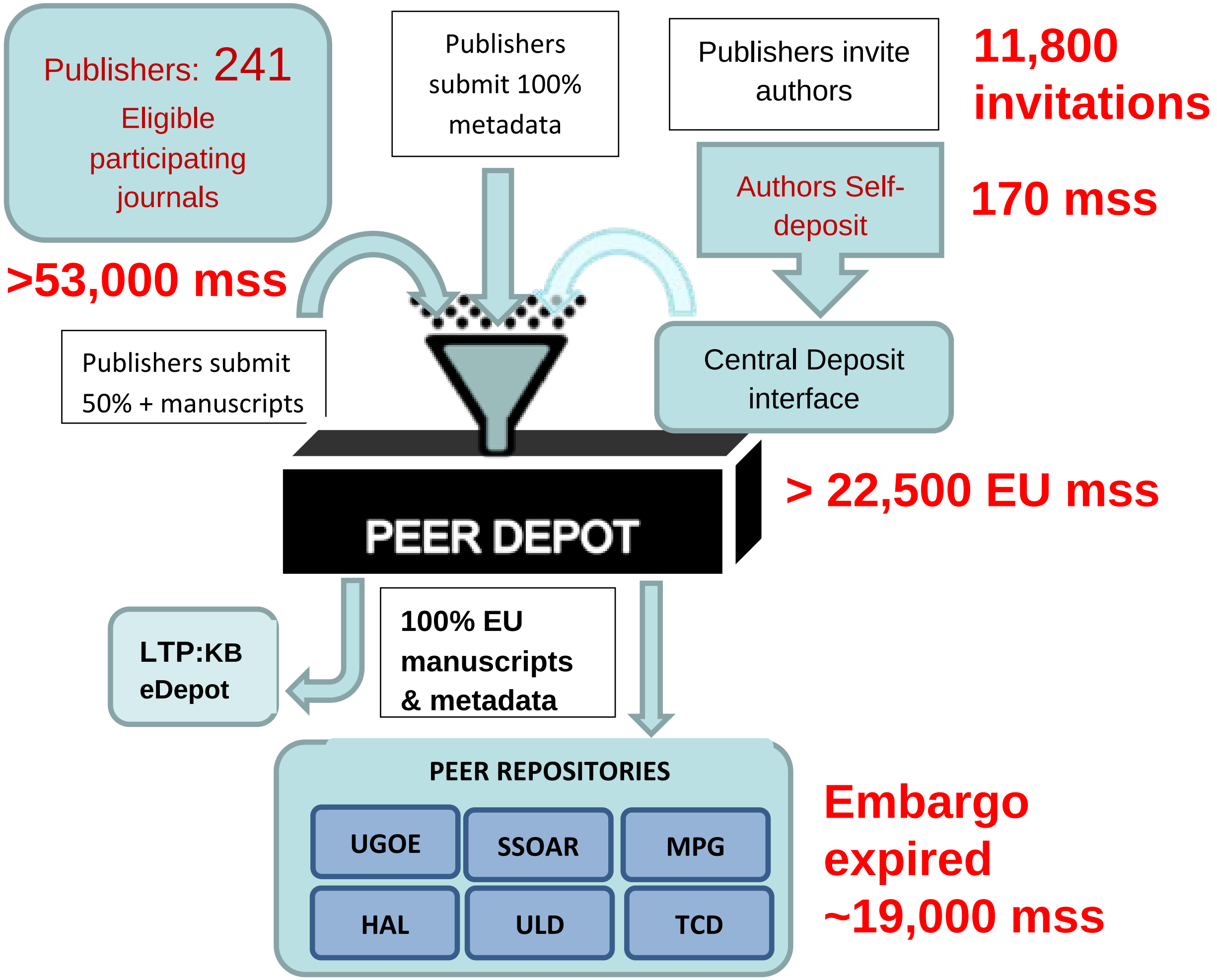
EU-Deposits processed (cumulated)



Author deposit 170

Embargo expired May 2012: ~19,000

The PEER Observatory – content levels



Publishers: 241

Eligible participating journals

>53,000 mss

Publishers submit 100% metadata

Publishers invite authors

11,800 invitations

Authors Self-deposit

170 mss

Publishers submit 50% + manuscripts

Central Deposit interface

PEER DEPOT

> 22,500 EU mss

LTP:KB eDepot

100% EU manuscripts & metadata

PEER REPOSITORIES

UGOE	SSOAR	MPG
HAL	ULD	TCD

Embargo expired ~19,000 mss

PEER Challenges and Solutions (1)

PUBLISHER CHALLENGES

•Stage two (accepted manuscripts) not standard extraction point

•Author accepted manuscripts in a variety of file formats

•All article types submitted

•Metadata delivery in several batches

- Article metadata are incomplete at acceptance time; Publication date unknown, DOI not attributed
- Extraction of only „EU“ authored manuscripts not possible at acceptance stage

•Different metadata formats

- NLM2.x, NLM 3.0, ScholarOne, proprietary

•Some Metadata elements delivered within PDF document

PUBLISHER / PEER DEPOT SOLUTIONS

➢Change Process at Publishers

➢Only one file format allowed – PDF

➢Checking mechanisms: journal/ article

- ➢ ISSN check
- ➢ article type check

➢Article kept until metadata completion

- ➢ Metadata are accepted in either one step (on publication) or two passes (on acceptance and on publication)
- ➢ EU author filter done at PEER Depot

➢Mapped onto single TEI structure

➢Extraction done at PEER Depot (GroBID) in order to increase content

PEER Challenges and Solutions (2)

REPOSITORY CHALLENGES

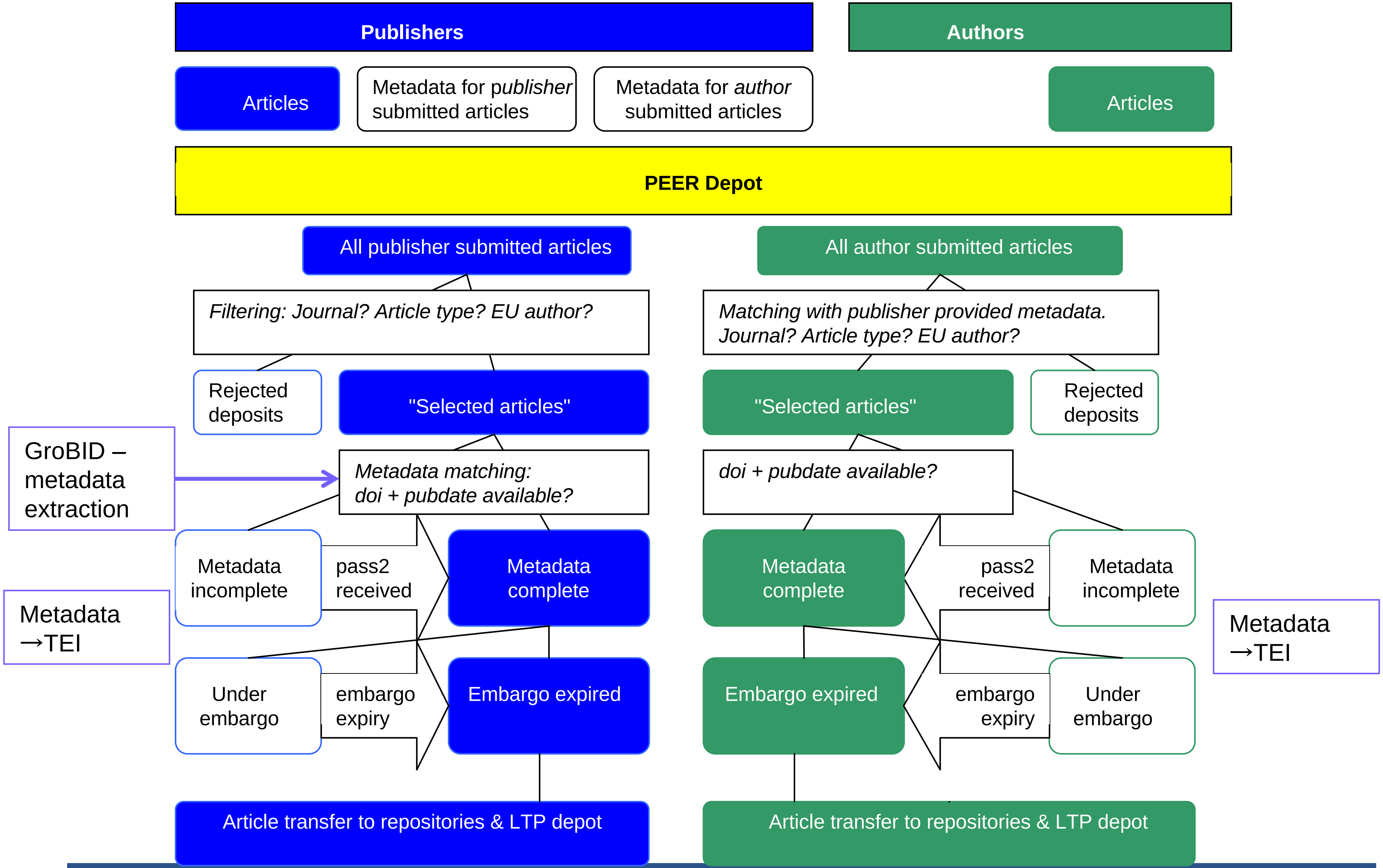
- **Varying metadata requirements**
- **Varying ingestion processes**
- **Hosting PEER content**
- **Not configured for accurate embargo management**
- **Author authentication**
- **Logfile provision**

REPOSITORY / PEER DEPOT SOLUTIONS

- **Convert TEI metadata into internally used metadata standard**
- **Implement SWORD protocol for transfer between Depot & repositories**
- **Build dedicated PEER Repository within framework of home institution**
- **Embargo management undertaken at PEER Depot (0 - 36 months)**
- **Central deposit interface at MPDL then transfer to PEER Depot**
- **Set up anonymisation process plus automated transfer to Usage team**

Other issues: Format and content problems with legacy manuscripts; Technical & financial challenges for repository participation (non PEER Partner repositories)

PEER Depot Workflow (what goes on in the black box)



Publishers

Authors

Articles

Metadata for *publisher* submitted articles

Metadata for *author* submitted articles

Articles

PEER Depot

All publisher submitted articles

All author submitted articles

Filtering: Journal? Article type? EU author?

Matching with publisher provided metadata. Journal? Article type? EU author?

Rejected deposits

"Selected articles"

"Selected articles"

Rejected deposits

GroBID – metadata extraction

Metadata matching: doi + pubdate available?

doi + pubdate available?

Metadata incomplete

pass2 received

Metadata complete

Metadata complete

pass2 received

Metadata incomplete

Metadata →TEI

Metadata →TEI

Under embargo

embargo expiry

Embargo expired

Embargo expired

embargo expiry

Under embargo

Article transfer to repositories & LTP depot

Article transfer to repositories & LTP depot

And when no meta-data is available

- **Automatic extraction of metadata from PDF**
  - – Typical use-case: IOP backfiles: Stage 3 documents used as input
- **Grobid: GeneRation Of BIbliographic Data**
  - – Machine learning environment for extracting metadata (and full text) from scholarly articles
    - Conditional Random Fields – plus some clever features
  - – Precise identification of
    - Authors (naming structure: first name(s), last name(s))
    - Affiliations (laboratory, department, institution)
    - Publication details (volume, issue, DOI)
    - Keywords, abstract
  - – Consolidation with CROSSREF
  - – Specific training per collection
  - – Generation of standardized TEI format
- **Reusable component (e.g. in publication repositories)**

PEER Observatory - Achievements

- **Enormous efforts made and results obtained**
  - Getting 6 heterogeneous repositories working in harmony on one project
  - Building the PEER Depot and creating infrastructural processes and protocols
  - Getting 12 very different publishers to contribute 241 test and over 200 control journals
  - Getting feeds for 241 heterogeneous journal systems to comply with PEER Depot requirements
  - Getting >53,000 mss processed the PEER Depot with uniform metadata
  - Ensuring that after EU filtering, each embargo group and subject has a statistically significant sample set of mss
  - Appointing and managing 3 leading research teams to work on the Observatory

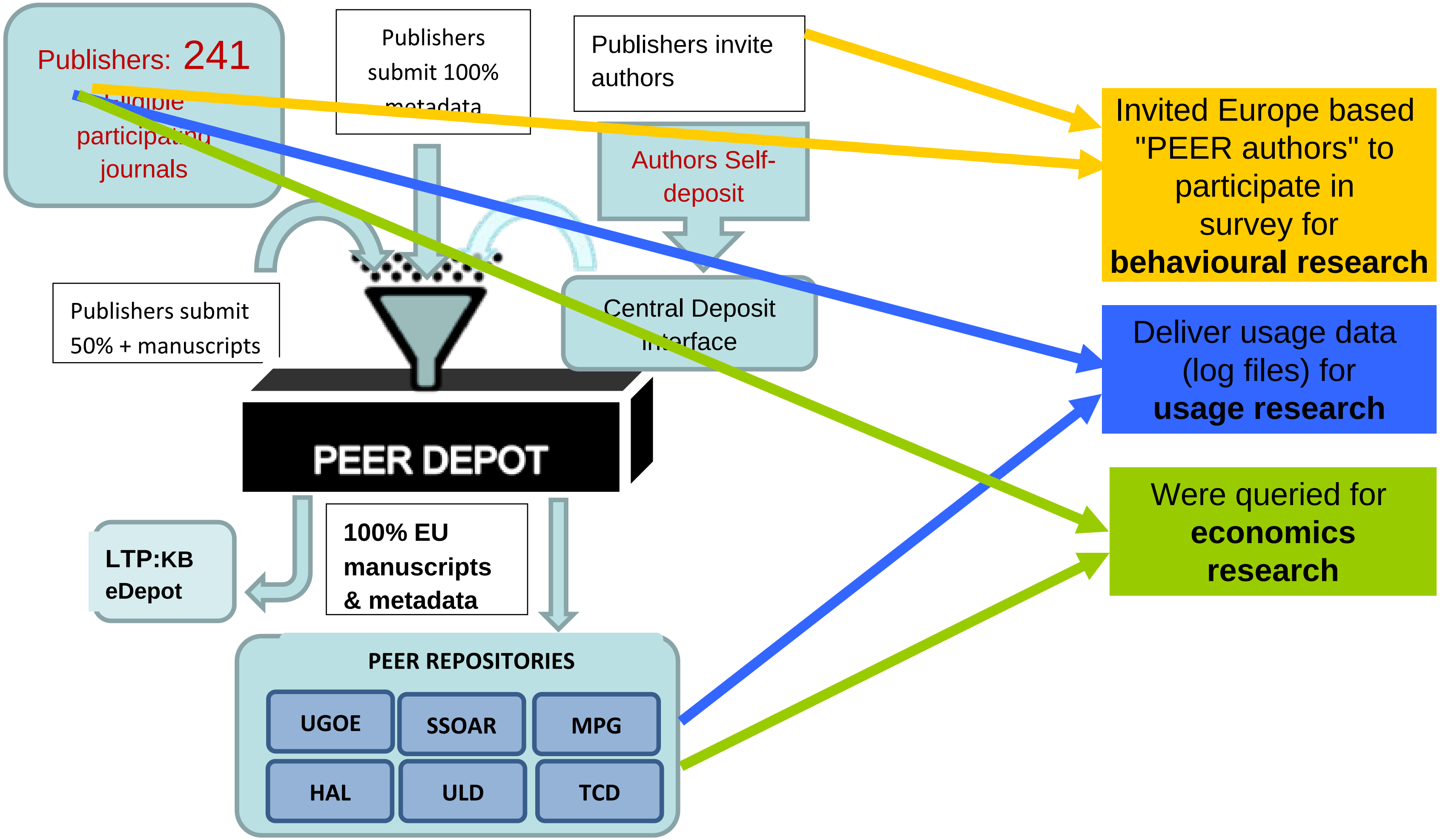
PEER Observatory - Achievements

- **Functioning collaborative infrastructure**
  - Linking repositories and publishers
  - Organising the transformation and flow of content
  - Metadata curation (quality control, embargo management etc.)
  - Usage data collected from repositories and publishers

- **Substantial quantities of content visible in repositories: ~19,000 EU deposits made publicly available (May 2012)**

  - **A working large-scale Observatory which has delivered results!**

The PEER Observatory & Research



Publishers: 241 Eligible participating journals

Publishers submit 100% metadata

Publishers invite authors

Authors Self-deposit

Publishers submit 50% + manuscripts

Central Deposit interface

PEER DEPOT

LTP:KB eDepot

100% EU manuscripts & metadata

PEER REPOSITORIES

UGOE	**SSOAR**	**MPG**
HAL	**ULD**	**TCD**

Invited Europe based "PEER authors" to participate in survey for **behavioural research**

Deliver usage data (log files) for **usage research**

Were queried for **economics research**