# DeepGreen

## Blending Data to Transform the German Scientific Publication Landscape to More Open Access

Thomas Dierkes
Julia A. Goltz-Fellgiebel

Cooperative Library Network Berlin-Brandenburg (KOBV)

ELAG 2018 Prague, 05. - 07. June 2018

# Short Project Profile

- Goal: Green open access compliant and automatic delivery of publishers' metadata & full texts to qualified repositories

- 1st funding period: 01.01.2016 - 31.12.2017
- 2nd funding period: 01.08.2018 - 31.07.2020 (approved in Mar '18)

- Funded by German Reseach Foundation (DFG)

- Project team
  - Kooperativer Bibliotheksverbund Berlin-Brandenburg (KOBV, project management)
  - Bibliotheksverbund Bayern (BVB)
  - Bayerische Staatsbibliothek (BSB)
  - Friedrich-Alexander-Universität Erlangen-Nürnberg, Universitätsbibliothek (FAU)
  - Technische Universität Berlin, Universitätsbibliothek (TU Berlin)
  - Helmholtz Open Science Koordinationsbüro am Deutschen GeoForschungZentrum (GFZ)

# Tasks to Solve

- We are living in a world of open access transformation!
- Alliance licences – i.e. national licence agreements co-funded by DFG –
  - . . . allow authors from authorised institutions for making their articles openly accessible via repositories
  - . . . are rarely used, partly because unknown, partly because too time consuming
  - . . . enables affiliated libraries to act accordingly but have little/no resources
- DeepGreen seeks to address these issues by some technical solutions
  - Which institutions are affiliated to a scientific article?
  - Does the journal belong to a licence agreement package for the given publication date?
  - Is the institution included as participant of the licence package?

# Issues to Consider

Make sure to . . .

- convince publishers to *continuously* deliver metadata (at least) & `.pdf` files
  - ✓ *S. Karger AG* and *SAGE Publications* actively support the project right from the beginning
- implement (all) interfaces as requested by publishers *and* repositories
- provide a database model for the licences at hand that is flexible enough

# Issues to Consider

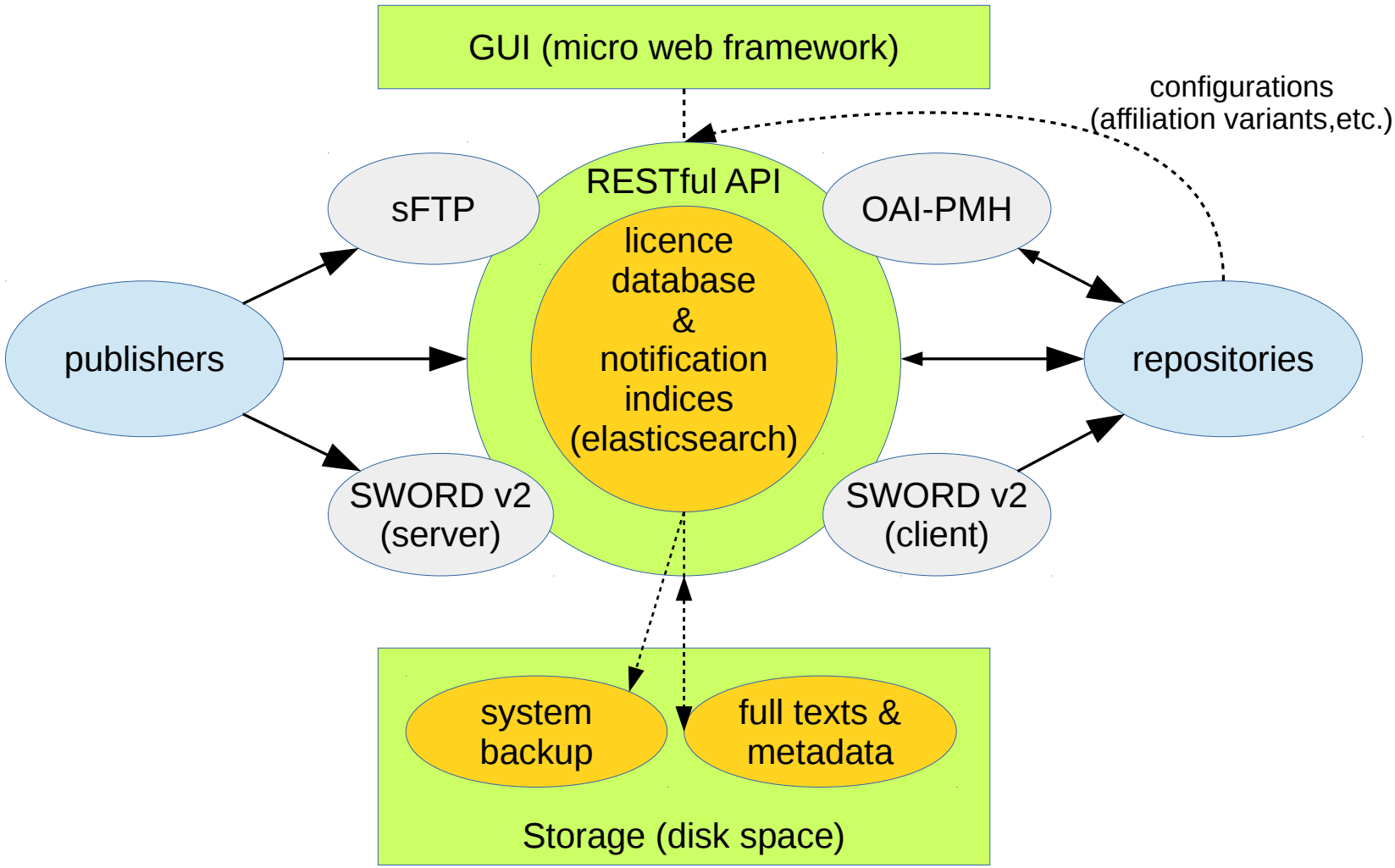
Make sure to . . .

- convince publishers to *continuously* deliver metadata (at least) & `.pdf` files
  - ✓ *S. Karger AG* and *SAGE Publications* actively support the project right from the beginning
- implement (all) interfaces as requested by publishers *and* repositories
- provide a database model for the licences at hand that is flexible enough

$\Longrightarrow$ Keep it Simple (Stupid)!

# DeepGreen System Architecture

based on Jisc Publications Event Router



# Technical Profile

- Based on (and refined) Jisc Publications Event Router
- Using *Flask v0.9* as mirco web framework
- NoSQL database engine *elasticsearch* (efficient & highly scalable)
- Import of `.xml` (DTD NLM/NISO JATS and **DTD RSC**) via
  - sFTP
  - SWORD v2
  - DeepGreen-API (RESTful)
- Export via OAI-PMH, SWORD v2, DeepGreen-API of (`.pdf` file and)
  - DTD NLM/NISO JATS and **DTD RSC**
  - Dublin Core
  - RIOXX
  - **METS/MODS**
  - **OPUS-XML**
  - **ESciDoc**
- Software: https://github.com/OA-DeepGreen
- Licence: Apache Licence, Version 2.0

# Matching and Mapping Algorithm

For each scientific article

1. **Check via [p|e]ISSN**:

   Is there a valid (alliance) licence applicable to the journal and the publication date?

   **No**: fetch next article, and start again

   **Yes**: analyse the affiliation field (and other matching criteria, e.g. grant IDs) to find all institutions of the article

   (*currently, this is all done by simple substring matches*)

2. **Check the affiliated institution(s) of the article**:

   For each, if included in the licence (of step 1.),

   a) **send** a hash key of the article ready for collection to the repository, or

   b) **push** the article to the respective repository (only if configured)

   c) **log** all of the matching and mapping for later controlling

3. **Log all non-matching as well** (!)

# Prototypical Implementation

DeepGreen ...

- is a **user-oriented service**, organised into accounts with profiles
  - publisher accounts with all import interfaces (since Oct '16)
  - repository accounts with all export interfaces (since May '17)
- is a **pure push-forward system** (no archive and no duplicate checks at all!)
- has a periodical link to
  *Elektronische Zeitschriftenbibliothek* (EZB), Regensburg (Germany)
  (*dynamical check of all journal licences packages*)
- can process metadata from publishers
  - **S. Karger AG**, **SAGE Publications**
  - **Europ. Math. Soc.**, **De Gruyter**, **Oxford Univ. Press**, **Royal Soc. Chemistry**
- delivers to repositories
  - **OPUS 4** (FAU), **DSpace** (TU Berlin), **ESciDoc/PubMan** (GFZ)

# Key Performance Indicators

| Publisher | Articles in 2015 | DeepGreen | % |
|---|---|---|---|
| S. Karger AG | 7 980 | 289 | 3.62 |
| SAGE Publications | 61 987 | 535 | 0.86 |

- Publishers provided articles of a whole publication year (2015)
- Numbers are obtained by using 248 standardised repository test accounts
  - all test accounts related to alliance licences of both publishers
- Outcome matches to 95% a manual sample of articles searched via `Scopus`, `WoS`, `PubMed` w.r.t. FAU, TUB and GFZ

# Perspectives for 2nd Funding Period 2018 - 2020

**Milestones to achieve** (not to be negotiable)

- Offer a legal framework for relationship between publishers and repositories
- Include a substantial number of publisher involved in alliance licences
- Consolidate the technical infrastructure
- Define and document a generic workflow for repositories

# Perspectives for 2nd Funding Period 2018 - 2020

**Milestones to achieve** (not to be negotiable)

- Offer a legal framework for relationship between publishers and repositories
- Include a substantial number of publisher involved in alliance licences
- Consolidate the technical infrastructure
- Define and document a generic workflow for repositories

$\Longrightarrow$ Start an operational service **DeepGreen** ($\beta$-phase) in 2019

# Perspectives for 2nd Funding Period 2018 - 2020

**Milestones to achieve** (not to be negotiable)

- Offer a legal framework for relationship between publishers and repositories
- Include a substantial number of publisher involved in alliance licences
- Consolidate the technical infrastructure
- Define and document a generic workflow for repositories

$\Longrightarrow$ Start an operational service **DeepGreen** ($\beta$-phase) in 2019

**To-do/wish list**

- Other licence types such as offsetting agreements or contracts of Research Information Services (FID)
  - Adaptation of DeepGreen to different legal settings
- Workflows suitable also for discipline-specific repositories
  - Subject-based classification/identification procedures
- Current research information systems (CRIS) as additional data recipients
  - Technical specification, interfaces, workflows

# DeepGreen Project Team

**KOBV / ZIB – Berlin**

- Prof. Dr. Thorsten Koch
- Beate Rusch
- Julia Alexandra Goltz-Fellgiebel
- Dr. Thomas Dierkes
- Jens Schwidder
- Laura Baumann (left Sep '17)

**BVB and BSB – Munich**

- Dr. Klaus Ceynowa
- Dr. Hildegard Schäffler
- Dr. Ortwin Guhling
- Michael Kassube
- Matthias Groß

**GFZ – Potsdam**

- Roland Bertelmann
- Heinz Pampel
- Kaja Scheliga
- Tobias Höhnow

**TUB – Berlin**

- Jürgen Christoph
- Monika Kuberek
- Per Broman
- Dagmar Schubert
- Pascal Becker
- Marsa Haoua (joined Apr '17)
- Melanie Janßen (joined Sep '17)

**FAU – Erlangen-Nuremberg**

- Konstanze Söllner
- Markus Putnings
- Cornelia Hoffmann
- Regina Heidrich

# Questions? Comments?

# Thank you for your attention!
## Imprint



**Email**: info@oa-deepgreen.de

**Contact**:

- Julia Alexandra Goltz-Fellgiebel (project management)
- Dr. Thomas Dierkes (software development)

**Project web page**: https://deepgreen.kobv.de/

**DeepGreen prototype**: https://www.oa-deepgreen.de/