# The ARCLib Project: An Open-Source Solution for Long-Term Preservation

Michal Růžička
Masaryk University, Institute of Computer Science
ELAG 2018
Prague, 2018-06-04

# High-Level Overview

- Research project funded by the Ministry of Culture of the Czech Republic within the NAKI program.
  - In the years 2016–2020.
  - 23 mil. CZK (~ 895,000 €).
- Cooperation of the Library of the Academy of Sciences of the Czech Republic, the National Library of the Czech Republic, the Moravian Library in Brno, and Masaryk University.
- **Development of open source solution for long-term archiving of digital documents (data) ARCLib.**

## https://arclib.cz/

# The Aims of the ARCLib Project

- Methodology for **long-term logical protection** of digital data.
- Methodology for **bit-level protection** and design of physical data storage.
- An **open-source** solution for long-term digital data archiving.
- **Verification in practice** by the pilot operation in a production environment.

# Project Team

- Library of the Czech Academy of Sciences (**LCAS**)
  - Martin Lhoták, Martin Duda, Jan Pokorský, Miroslav Pavelka, Pavel Madar, Ivana Šlapáková, Martina Nezbedová, Andrea Fojtů-Miranda, Jan Hutař, Jana Křížová
- Masaryk University (**MUNI**)
  - Miroslav Bartošek, Vlastimil Krejčíř, Michal Růžička, Lukáš Hejtmánek
- National Library of the Czech Republic (**NL**)
  - Zdeněk Vašek, Václav Jiroušek, Iveta Lodrová, Natálie Ostráková
- Moravian Library in Brno (**MLB**)
  - Petr Žabička, Zdeněk Hruška

# Initial State

- The **National Digital Library** project (NL, MLB).
- The **Czech Digital Library** project and **Kramerius**, **ProArc**, and **RDflow** tools (LCAS, NL).
- LTP Pilot Project – **testing of Archivematica** (MLB, MUNI).
- Foreign open-source projects – Archivematica, **RODA**, ...
- Commercial solutions – **Rosetta**, **Tessella**, ...

# Distribution of Tasks

- Preparation of methodology for logical data protection – **LCAS**
- Preparation of methodology for bit-level protection of data – **MUNI**
- Integration of Archivematica and DSpace into ARCLib – **MUNI**
- Integration of ProArc and Kramerius systems into ARCLib – **LCAS**
- Defining standards for exchangeable information packages within ARCLib – **NL**
- Coordination of works on ARCLib – **LCAS**

# Methodology for Logical Data Protection

- **General / theoretical part.**
  - The first part covers **the scope of long-term preservation (LTP)**, a description of the **basic standards** used in the field, and **the usual strategies and practices**.
- **Practical part linked to ARCLib.**
  - The second part of the methodology documents the process of **the practical implementation of the requirements** of the long-term preservation solution, including **suggestions of specific SW** solutions.
- **Implementation section.**
  - The third part of the methodology contains **recommendations** for ARCLib users for **each type of data** and will describe the standards set during the pilot operations in the production environment.

# Methodology for Bit-Level Protection of Data

- Research on possibilities and technologies for **efficient and secure physical storage of large volumes of data**.
- The LTP system should be built as highly modular, and **the archive storage should be one of the modules** with a **clearly defined communication interface**.
- To ensure the physical security of the repository, it is necessary to **distribute** physical **copies of data** to more **geographically separate locations**.
- Handling **the high demands** on available **capacity** and system **throughput** in **massive parallel workloads**.
- Comparison of existing distributed storage technologies such as **GlusterFS**, **dCache**, **Luster**, **HDFS**, **Ceph**, and so forth, or the implementation of a custom solution using **Btrfs**/**ZFS** file systems.

# Integration of Archivematica and DSpace into ARCLib

- **Testing** of Archivematica (AM).
  - Processing of **mid-size** (up to tens of GiB) **data packages**.
  - Improvements in the processing of **large volumes of small-size packages**.
  - AM developments are likely to continue to optimize and increase performance (**parallel processing**).
  - AM is **still hard-to-configure** in many cases and suffers from **instability**.
- **Tuning** of AM **workflow to receive AIPs** exported **from DSpace**.

# Integration of ProArc and Kramerius Systems into ARCLib

- **ProArc export** formats:
  - Kramerius.
  - NDL (National Digital Library standard).
  - Full ProArc XML (FOXML)
- **Kramerius**
  - Will be enhanced with **data converter for ProArc**.

# Defining Standards for Exchangeable Information Packages within ARCLib

- **Interoperability** with the LTP system of the National Library of the Czech Republic.
  - **NDL packages** are one of the basic inputs to ARCLib.
- **Analysis** of different **types** of **input data**.
  - NDL packages.
  - ProArc packages.
  - DSpace – AM packages.
- 2016 design of ARCLib AIP.
  - **SIP = DIP**.
  - **ARCLib AIP XML** with metadata.

# ARCLib AIP

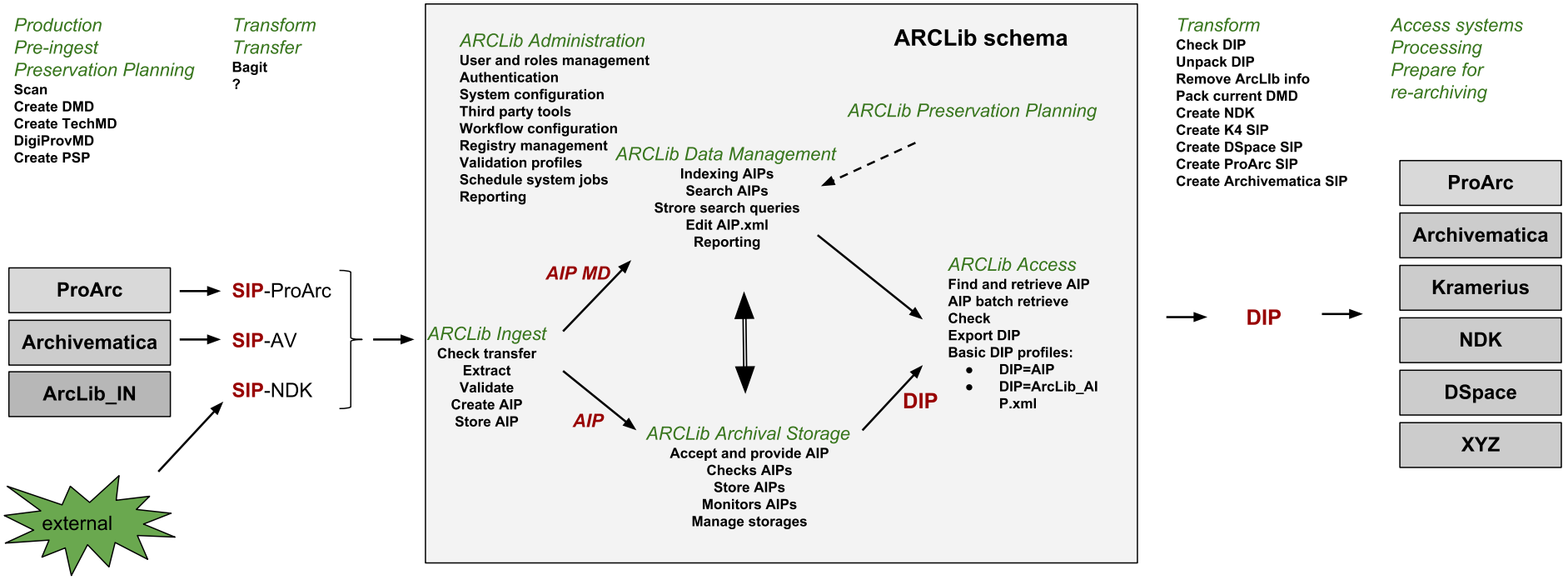
ARCLib AIP consists of two parts:

1. **SIP** from the data provider.
    - NDL, ProArc, AM.
2. **ARCLib AIP XML** metadata.
    - Partially generated from SIP
        - **Bibliographic metadata** – Dublin Core + MODS
        - **Technical metadata** – scanner type, date of scanning, operator, JHOVE data, …
    - and data generated by ARCLib from received SIP.
        - **Administrative metadata** – data provider, workflow, validation log, validation profile, identification of identity formats, date, …

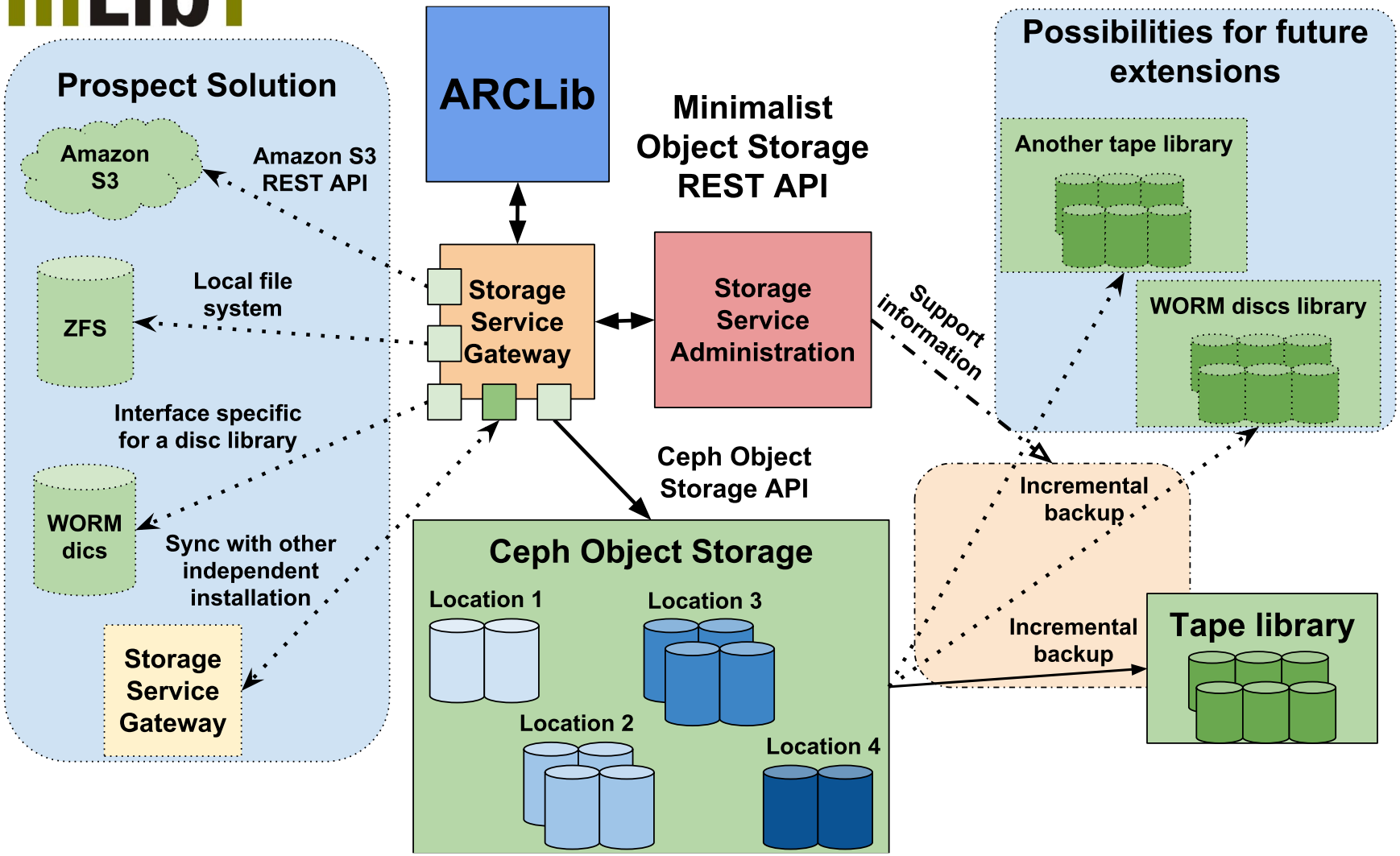
# ARCLib System

- Implementation of **OAIS functional models** (ČSN ISO 1471).
- **Not** going to **reimplement functionality provided** by ProArc or Archivematica.
  - The tools are used to create SIP packages.
- ARCLib **developments focus** on the creation of **essential modules**:
  - ARCLib Ingest.
  - ARCLib Data Management.
  - ARCLib Archival Storage.
  - ARCLib Administration.
  - ARCLib Access.

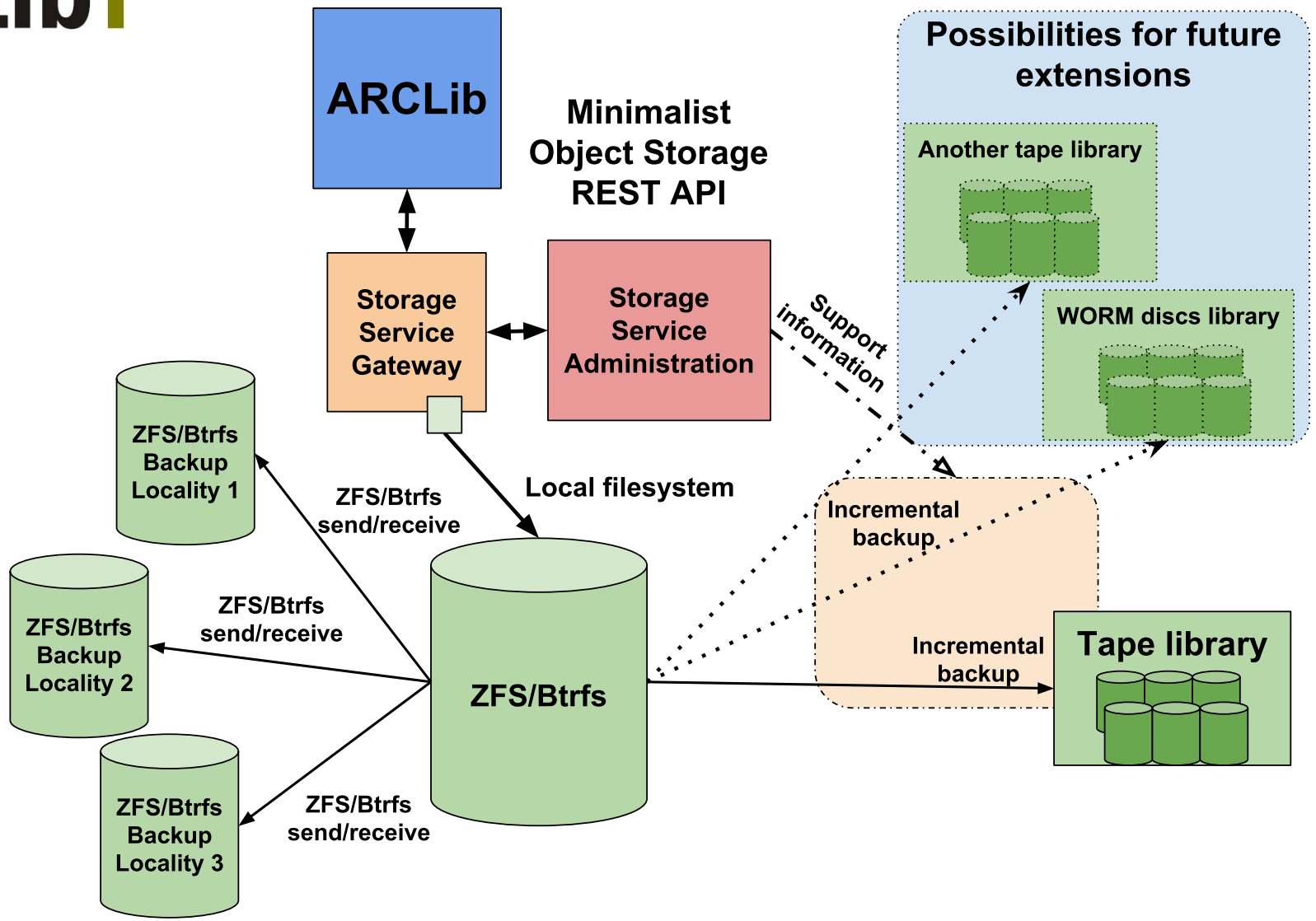
# ARCLib High-Level Schema



# ARCLib Archival Storage



# ARCLib Archival Storage



# Project Schedule

- 2016 – Architecture design.
- 2017 – Methodology for logical data protection.
  – The launch of programming, prototypes.
- 2018 – Methodology for bit-level data protection.
- 2020 – ARCLib LTP system v1.0.
  – Verification of the archiving solution ARCLib in the pilot operation in the production environment in the Library of the Czech Academy of Sciences.

# Methodology for Logical Data Protection

**Author's team:**

- Eliška Pavlásková
- Zdeněk Vašek
- Jan Hutař
- Andrea Miranda
- Zdeněk Hruška

84 pages, **certified by the Ministry of Culture** of the Czech Republic.

# Methodology for Logical Data Protection

**The theoretical part:**

- Describes the general procedure.
- Designed for all types of storage.
- Policies for building a trusted long-term storage.
- The concept of OAIS functional units.
- Long-term storage strategy.
- Recommendations for DSA certification.

# Methodology for Logical Data Protection

**Practical and implementation part:**

- Tied to the ARCLib solution.
- System architecture description.
- ARCLib AIP XML – AIP metadata specification, ARCLib information package solution using common standards.
- Organizational and personnel operational recommendations.
- Recommendations for financial planning and external tools/services.

# Methodology for Logical Data Protection

**Areas of application:**

- Institutions seeking long-term digital data retention solution.
- Users of the ARCLib system.
  - Recommendations on OAIS-compliant system creation.
  - Not only technical solutions,
  - but also organizational structure, procedures, and processes…
- Material for university education – readily available summary text.


**Thanks for the attention**

Michal Růžička
ruzicka@ics.muni.cz
https://arclib.cz/