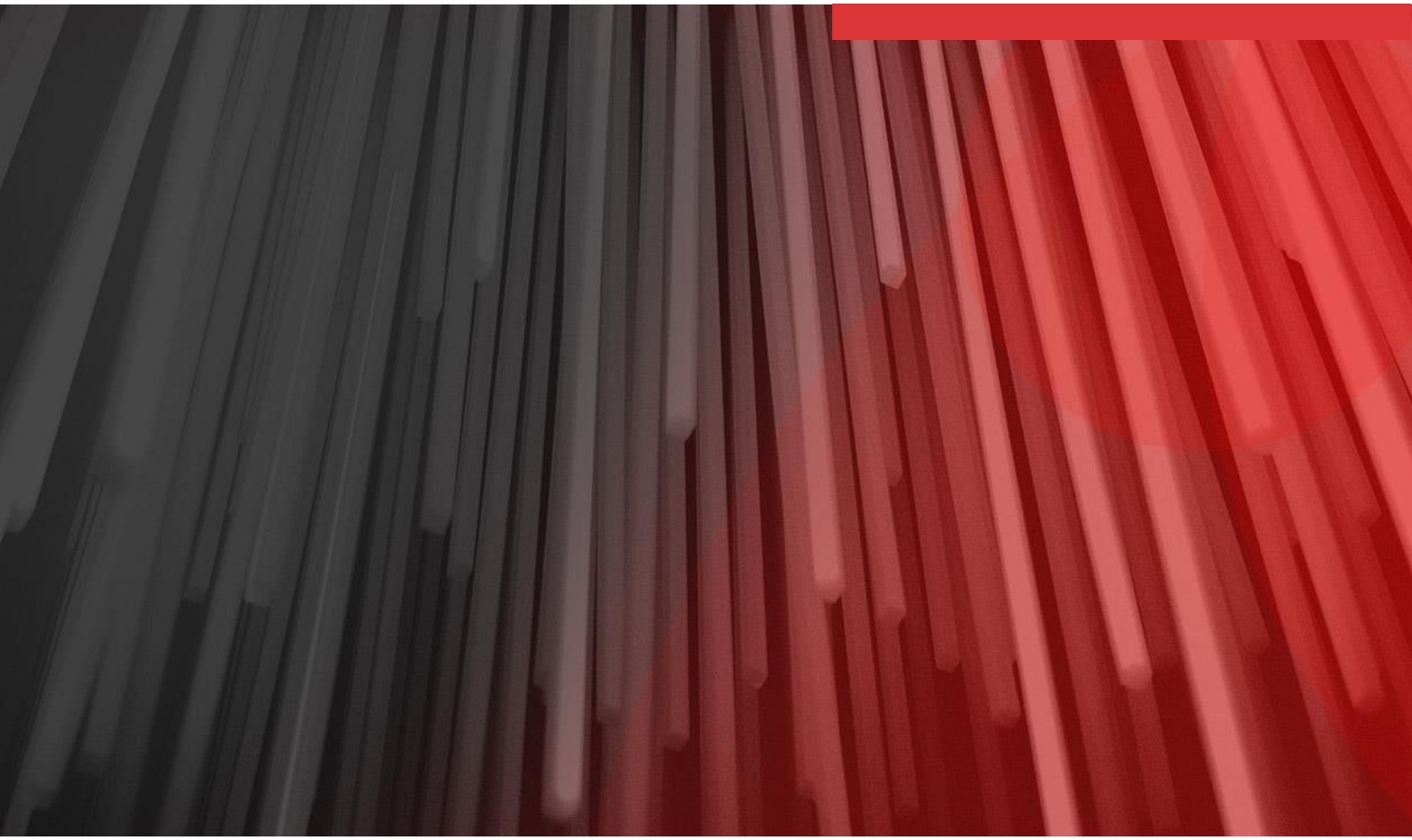
keep.
Preserving the future

RODA 3



LONG-TERM DIGITAL PRESERVATION

CHARACTERISTICS AND TECHNICAL REQUIREMENTS

ABOUT THIS DOCUMENT

Identifier	WP181216		
Approved by	Luís Miguel Ferros	**Approved on**	2018-11-29
Classification	Public		
Distribution	N/A		

REVISIONS

#	Date	Authors	Modifications
1	2018-11-29	Miguel Ferreira	First version of the document

EXECUTIVE SUMMARY

RODA is an open-source long-term digital preservation system that delivers functionality for all the main functional units of the OAIS reference model. RODA is capable of ingesting, managing and providing access to various types of digital content produced by large corporations and public bodies.

RODA was developed using open-source technologies and it is supported by standards such as the Open Archival Information System (OAIS), Metadata Encoding and Transmission Standard (METS), Encoded Archival Description (EAD), Dublin Core (DC), E-ARK Information Package specifications and PREMIS (Preservation Metadata).

This document describes the main features and value propositions associated with this software. The document also outlines the technical requirements and what is necessary to deploy the software in a production environment.

PRESERVE AND PROVIDE ACCESS TO DIGITAL MATERIAL PRODUCED BY LARGE ORGANISATIONS

RODA

RODA (Repository of Authentic Digital Records) is a long-term digital repository solution that delivers functionality for all the main functional units of the OAIS reference model. RODA is capable of ingesting, managing and providing access to various types of digital content produced by large corporations and public bodies.

RODA was developed using open-source technologies and it is supported by standards such as the Open Archival Information System (OAIS), Metadata Encoding and Transmission Standard (METS), Encoded Archival Description (EAD), Dublin Core (DC), E-ARK Information Package specifications and PREMIS (Preservation Metadata).

It implements an ingest workflow that not only validates standardised SIPs, but also checks its content for virus, handles file format identification, extracts technical metadata, and migrates file formats to more “preservable” alternatives.

RODA also provides access to digital information in several forms, such as searching and browsing via graphical user interfaces, as well as providing REST APIs for systems integration. Discovery services are provided over both descriptive metadata and textual content (automatically extracted from an array of document-based formats). Online consultation of ingested objects, preservation formats and dissemination derivatives are also possible via the repository catalogue.

Administration interfaces allow repository managers to maintain representation information, to manage repository risks and to execute preservation actions on all object types available in the repository.

RODA ensures that ingested data remains authentic by recording PREMIS metadata every time an action is performed on a digital object. It records provenance information in archival metadata standards such as EAD or DC and ensures integrity and availability by frequently monitoring data and making sure that it has not been tampered with. All interactions between users and the repository (human and software) are logged for security and accountability reasons.

Conforms to open standards

RODA is compliant with several open descriptive metadata standards such as EAD 2002, EAD 3 and DC, PREMIS for preservation metadata and METS for structural metadata.

It also has the ability to support more standards using an advanced templating system (to support searching, viewing and editing of metadata).

SIP, AIP and DIP formats are also based on open specifications compliant with various repository implementations to avoid technology lock-in.

Vendor independent

RODA is 100% built on top of open-source technologies.

The entire infrastructure required to support RODA is vendor independent. This means that you may use the hardware and the Linux distribution that best fit your institutional needs.

Because the product itself is open source, you don't have to rely on a single vendor for support.

Authenticity

RODA uses preservation metadata (PREMIS) to create a trust chain between all generations of data.

Preservation metadata, together with the establishment of trust of its surrounding environment (ISO 16363) ensures that the service is reliable and that the enclosed digital records are authentic.

RODA also comes with plugins that assess the validity of digital signatures during ingest and has the ability to re-sign archived documents when the lifetime of digital signatures is coming to an end.

Support for multiple formats

RODA is capable of ingesting all sorts of content. Migration action components are embedded in the system for coping with decaying text documents, raster images, relational databases, video, and audio by normalizing them to formats more adequate for long-term preservation.

A task execution engine and a plugin system enable RODA to easily support additional format migrations.

Additionally, representation information networks can be managed within the repository itself, letting you opt for the right preservation strategy at the right time.

Advanced **ingest workflow**

RODA supports the ingest of digital material as well as any associated metadata in several distinct formats. Tools are provided to enable Producers to create packages in the supported Submission Information Package formats (SIP). The ingest workflow can be customised by the user in order to implement institutional policies and handle special collections of data.

PRESERVATION STRATEGIES SUPPORTED BY RODA

RODA was designed to be flexible enough to cope with every preservation strategy found on the literature. It can natively support format migration and format normalization, encapsulation and provide support for emulation (not included off-the-self).

FORMAT MIGRATION

Format migration consists of the conversion of resources from one file format to another (e.g., conversion of Microsoft Word to PDF or to OpenDocument) with the purpose of making them more accessible to their designated community or to convert them to formats that, given their intrinsic characteristics, are considered more adequate for long-term preservation[1].

RODA natively supports the conversion of hundreds of file formats via its task execution engine and plugin system. The extensible nature of RODA enables it to be updated at any time to cope with new file formats and to support more advanced preservation tasks.

Quality assurance and preservation metadata ensure records remain authentic while providing traceable records of all changes and events that occur to a digital representation. All actions performed by users and software are logged for security and accountability reasons.

ENCAPSULATION

Encapsulation is a method based on the premise that preserved objects should be self-describing, linking content with all of the information required for it to be deciphered and understood whenever it is needed. Representation Information is a key concept in this context. Representation Information is the information that maps a data object into more meaningful concepts. An example is the ASCII definition that describes how a sequence of bits (i.e., a data object) is mapped into a symbol. In summary, "data interpreted using its Representation Information yields information". Representation Information can be divided into three classes:

- Structural information, describes the format and data structure concepts to be applied to the bit-stream, which result in more meaningful values like characters or number of pixels.
- Semantic information, this is needed on top of the structure information. If the digital object is interpreted by the structure information as a sequence of text characters, the semantic information should include details of which language is being expressed.

[1] https://en.wikipedia.org/wiki/Digital_preservation

- **Other Representation Information**, includes information about relevant software, hardware and storage media, encryption or compression algorithms, and printed documentation.

Representation Information is natively supported by RODA. Digital objects can be linked to Representation Information records, which can reference other Representation Information records, thus creating a network of Representation Information records. This is called a Representation network. The entire representation information network can be managed via the repository user interfaces, including the creation of links between archived files (and/or representations) and their relevant representation information records.

This strategy is particularly important in the context of preserving research data where no standard formats are typically available and there is a high demand for information that facilities the understanding of content, rather than formats.

EMULATION

Emulation consists of replicate the functionality of an obsolete system using current technologies. According to van der Hoeven, "Emulation does not focus on the digital object, but on the hardware and software environment in which the object is rendered. It aims at recreating the environment in which the digital object was originally created."[2].

RODA is able to retain the original versions of digital representations as they have been received via the ingest process. These original versions are stored inside Archival Information Packages (AIP) and are an intrinsic part of the overall preservation process.

Keeping the originals inside the AIP means that emulation strategies remain viable in the given future. The implementation of emulation techniques during access means that formats that cannot be preserved any other way can now be delivered to consumers with no degradation.

[2] https://en.wikipedia.org/wiki/Digital_preservation

Embedded preservation actions

Preservation actions can be executed right from the user interface over any selection of digital objects in the repository.

The task execution engine enables the repository to parallelise the task execution process in order to take full advantage of the existing processing power.

Preservation actions include format conversions, checksum verifications, virus checks, various maintenance tasks, risk assessment, etc.

Scalable

The service-oriented nature of RODA allows it to be highly scalable, enabling the distribution of processing load between several servers.

The use of advanced indexing and parallelisation frameworks enables RODA's discovery services to be spread across multiple servers for greater performance and to take advantage of all the CPUs available in each server to mass process thousands of objects at the same time.

Copes with the rapid changing nature of technology

The pluggable architecture of RODA makes it easy to add more functionality to the system without affecting the core.

This includes adding new preservation tasks such as preservation actions, risk assessment tools, internal and external monitoring, etc.

The system also manages data in a well-documented open Archival Information Package (AIP) structure that can be easily inspected by users and ingested by other repository systems. This way, your data is never imprisoned inside a single system.

Advanced access control

Users must be authenticated before accessing any functionality and objects in the repository. All user actions are logged for future accountability.

Permissions are fine-grained and can be defined at the top of the repository level, all the way down to individual data objects.

Authentication is supported by a Central Authentication Service (CAS) that is able to connect to various authentication services such as LDAP, Active Directory, OAuth 1.0/2.0, custom database, OpenID, RADIUS, SPNEGO (Windows), Trusted remote user, X.509 (client SSL certificate), etc.

Integration with 3rd party systems

RODA exposes all its functionality via well-documented REST API. Convenient Java libraries are available on GitHub to allow developers to interact with RODA via its Core APIs. Several tools exist to create and manipulate SIPs and submit them to RODA's ingest workflow. In fact, RODA ability to ingest data from other document management systems include: 1) data typically available on the filesystem via RODA-in tool, 2) data stored on relational databases via the Database Preservation Toolkit, and 3) via direct connectors to original systems APIs.

RISK ASSESSMENT AS THE APPROACH TO PRESERVATION MONITORING AND PLANNING

Digital preservation is often defined as a risk management exercise where the aim is to convert the uncertainty about maintaining the usability of authentic digital objects into quantifiable risks. The purpose of a digital repository is to do everything it can to mitigate the risks that prevents its ability to provide the access to authentic digital information. The measure of success of a repository's work is the "quality" of information it releases to its users[3].

It is easy to see that risks are not only technological but also organizational, staff and systems-related, and connected with the external factors arising from the environment where the digital repository operates. Like any organization, digital repositories can benefit from risk analysis and risk management techniques to support both their general management and their core business of digital curation and preservation.

RODA development team agrees that following a risk management methodology is a good approach to digital preservation and decided to embed it into the repository's preservation workflow.

RODA is flexible in what concerns the types of data to be ingested. The ingest workflow can be customised to support all sorts of institutional policies ranging from the very strict, where only a very limited number of file formats and metadata standards are accepted by the repository, to the very loose one, where the repository accepts anything that comes along the ingest pipe.

Taking advantage of a risk management approach, the preservation manager can decide after ingest what is the best action to take to preserve the archived data, and not start immediately to create artificial constraints to the intake of new data.

Planning based on **risk assessment**

The preservation manager is expected to do risk analysis on the archived data and plan adequate mitigation strategies after assessing the level of risk that the repository is undertaking. Risk assessment tasks are implemented as plugins in the repository system and typically include responsibilities such as detecting files for which the repository does not maintain enough representation information, or files that have suffered tampering, or files being affected by bit-rot, or files whose file format is being replaced by a new open-standard, or files whose designated community has difficulties in reading, etc.

After a proper risk assessment, the preservation manager can decide whether he/she wants to accept the risk or mitigate it via a risk mitigation plugin.

[3] Most of the text found in this section is a verbatim copy with small adaptations of the *Digital Repository Audit Method Based on Risk Assessment* (DRAMBORA) published by the Digital Curation Centre (DCC) and the DigitalPreservationEurope (DPE).

WHY OPEN-SOURCE IS IMPORTANT IN THE CONTEXT OF DIGITAL PRESERVATION

The open-source software movement represents a forty-year-old software development and a distribution philosophy that offers several valuable advantages to the digital curator. While these advantages can be identified in a wide range of application areas, there are several intrinsic qualities that lend themselves particularly well to digital preservation, and that make the use of open-source an excellent starting point for data creators, curators and re-users seeking to facilitate the long-term use of digital materials[4].

Several open-source applications are among the most proven and reliable digital solutions available in the market (take MySQL, Apache Web server and Firefox as examples). By regularly embracing the concept of open standards, these technologies further remove the mystery from information storage over the longer term making it more transparent and accessible for its users. As with source code disclosure, open standards aim to excise the opaque veneer that threatens and disrupts digital preservation, limits and shortens access to long-term stored materials, and hampers the straightforward interchange of digital content.

The adoption of open-source software provides several benefits throughout the entire scope of the digital curation life-cycle. To determine the sustainability of an application or file format, several important criteria must be considered. These include its longevity, the ease of its re-creation or emulation, its adherence to and use of open standards, the level of legal freedom associated with its use, its associated costs, its ubiquity, its support for metadata and its stability.

From the very conception of digital information, open-source presents some immediate advantages. With open-source software, acquisition costs are certainly lower than those associated with equivalent proprietary products, and although some costs are involved in implementing and maintaining an open-source infrastructure, several studies agree that total costs of ownership are significantly cheaper than those of proprietary software. Furthermore, transparency through source-code availability and the frequent association between open-source and open standards facilitates long-term comprehension and re-use, enabling creators, curators and re-users to effectively and explicitly present their digital materials alongside their underlying descriptive infrastructures.

In addition, having the source-code of the application freely available on the Web has the greater advantage of preventing situations of vendor lock-in. A customer that becomes dependent on a vendor for its products and services is unable to switch to another vendor without entering in substantial investments. Open-source circumvents this situation by enabling any vendor to

[4] Most of the text found in this section is a verbatim copy with small adaptations of the report published by the Digital Curation Centre with the title "Instalment on Open Source for Digital Curation" written by Andrew McHugh.

provide services on top of the same product with equal chances of delivering a quality service to its customers.

TECHNICAL REQUIREMENTS

RODA requires a minimum of two computers to operate: a server and a workstation. The server is responsible for hosting the data and handle all business processes. The workstation is used by end-users to administer the system.

The following sections outline the minimum requirements to run the application on both of these computers.

SERVER

RAM	4 GB 8 GB recommended
CPU	1.0 GHz Quad-Core or superior
HDD	20 GB Depends of the volume of data to be ingested.
Operating system	Ubuntu Server 16.04 LTS or compatible No licensing costs
Software	Docker engine
Network	100 Mbit/s or superior 1 Gbit/s recommended

NOTE: Please note that real-life production environments are usually more complex, often requiring a cluster of server machines in order to meet highly-demanding performance requirements of large preservation services.

WORKSTATION

RAM	4 GB
CPU	Intel Dual-Core or superior
Screen	1280x768 pixels or superior
Operating system	Windows/Linux/MacOS
Software	Web browser
Network	100 Mbit/s or superior 1 Gbit/s recommended

KEEP SOLUTIONS

KEEP SOLUTIONS is a company whose mission is to provide advanced solutions for information management and digital preservation.

Our approach consists in providing software and services to allow our customers to make a more efficient management of their information assets.

The company started its activity in 2008, having acquired the status of academic spin-off of the University of Minho, for being a business initiative with strong bonds with research centres and departments from this institution.

Our clients are mostly found in the public sector, more specifically in the areas related to archives, libraries and museums.

We invest in the continuous development of innovative solutions. To support that, we remain active in the production of scientific knowledge while engaging in large-scale R&D projects in cooperation with national and international institutions.

www.keep.pt
+351 253 066 735
info@keep.pt
sales@keep.pt
KEEP SOLUTIONS, LDA.
Rua Rosalvo de Almeida, nº 5,
4710-429 Braga
Portugal