Digital Preservation and AI - Critical Challenges

Dr. **Hrvoje Stančić**, full prof.
Vice dean for organization and development /
Chair of archival and documentation sciences
Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences
University of Zagreb, Croatia
hstancic@ffzg.hr

Contents

1. Introduction
2. AI or automation?
3. Research
4. Research – aim and methodology
5. Research results
6. Conclusion

1. Introduction

- Archival institutions – information society challenges
- Emerging technologies
  - change information landscape
  - new user habits and expectations
  - redesign of the relationships between users and institutions
  - traditional practices of archiving are being transformed
- Disruptive technologies
  - artificial intelligence, blockchain, big data, crowdsourcing, gamification, etc.
  - positive disruption of current archival processes (service improvement)

1. Introduction …

- Requirements for the (long-term) preservation (LTP) of digital resources in light of constant change and development of ICT
  - LTP actions = conversion, migration, emulation, virtualization

- LTP challenges – how to preserve
  - authenticity
  - integrity
  - reliability
  - usability
  - non-repudiation
  - security
  - confidentiality
  - proof of ownership

⇒ Trustworthy records

- authentic, accurate, reliable

2. AI or automation?

- "**Automation** saves time and money spent on monotonous, voluminous tasks and gives employees an opportunity to apply themselves to more complex processes."
- "**AI** deals with technologies, systems or even processes that competently mimic how human beings make decisions, react to new information, speak, hear, as well as understand language."

Mark Nasila

https://www.coriniumintelligence.com/insights/artificial-intelligence-vs-automation

- Intelligent automation?

2. AI or automation? …

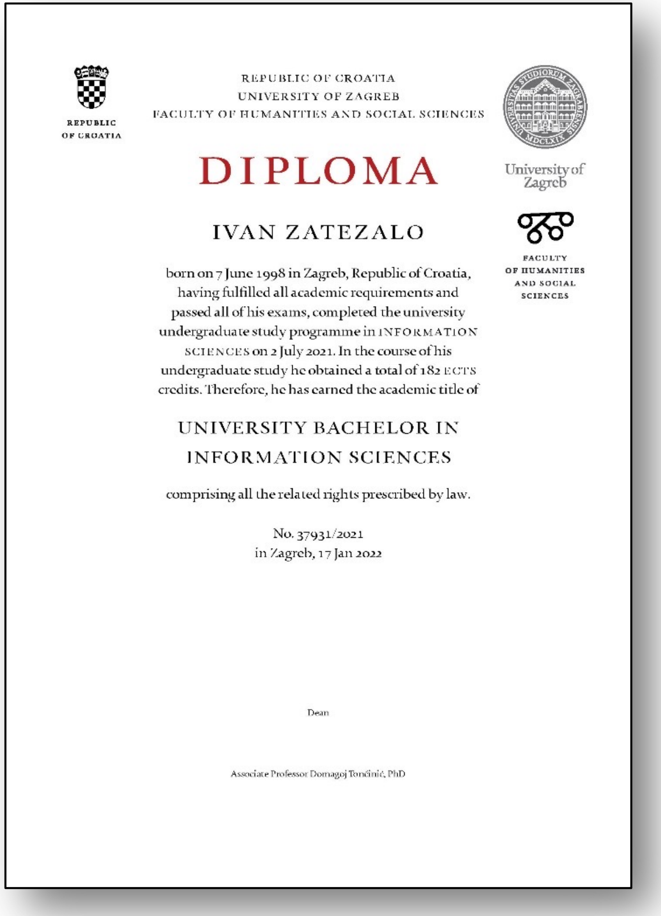
- Robotic Process Automation (RPA)
  - software robots or "bots" (similar to, but more advanced than macros in e.g. Word)
  - automation of series of tasks by mimicking human interaction with (different) software solutions
  - business process automation
  - eliminates high-volume, rule-based repetitive tasks
  - can be combined with other disruptive technologies, e.g. blockchain

2. AI or automation? …

- RPA – example
  - Faculty of Humanities and Social Sciences (FHSS), University of Zagreb – **Blockchain-based diploma authentication system**
  - Motivation for starting the project
    - FHSS annually issues around 1,300 diplomas
    - forged diplomas “issued” by the FHSS have been found by employees
    - requests for authentication of diplomas
      - in one year, we have checked authenticity of 5,500 FHSS diplomas and additional 7,500 diplomas were requested to be authenticated although other faculties issued them (requests wrongly addressed to FHSS)
    - if only 5 mins needed for verification of 1 diploma =
      27 weeks for 1 employee (1/2-year FTE)

2. AI or automation? …

Scanning QR code from the diploma back



REPUBLIC OF CROATIA

REPUBLIC OF CROATIA
UNIVERSITY OF ZAGREB
FACULTY OF HUMANITIES AND SOCIAL SCIENCES

University of Zagreb

FACULTY OF HUMANITIES AND SOCIAL SCIENCES

DIPLOMA

IVAN ZATEZALO

born on 7 June 1998 in Zagreb, Republic of Croatia, having fulfilled all academic requirements and passed all of his exams, completed the university undergraduate study programme in INFORMATION SCIENCES on 2 July 2021. In the course of his undergraduate study he obtained a total of 182 ECTS credits. Therefore, he has earned the academic title of

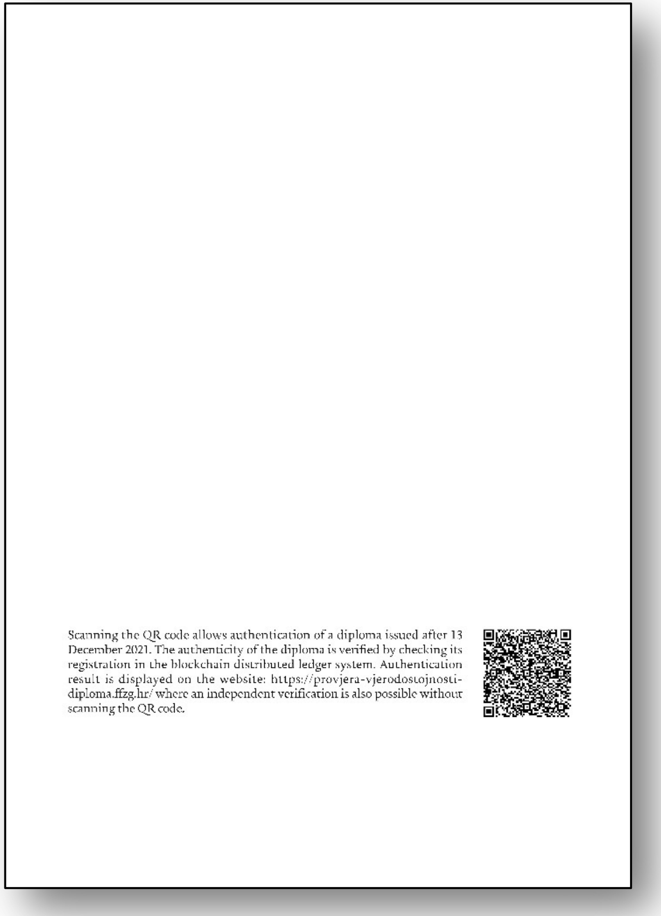
UNIVERSITY BACHELOR IN INFORMATION SCIENCES

comprising all the related rights prescribed by law.

No. 37931/2021
in Zagreb, 17 Jan 2022

Dean

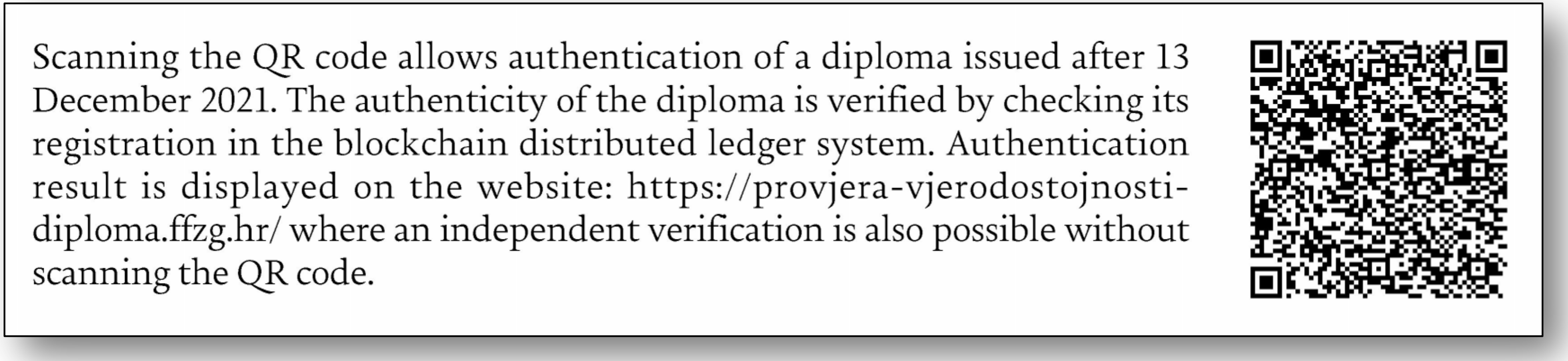
Associate Professor Domagoj Tončinić, PhD



Scanning the QR code allows authentication of a diploma issued after 13 December 2021. The authenticity of the diploma is verified by checking its registration in the blockchain distributed ledger system. Authentication result is displayed on the website: https://provjera-vjerodostojnosti-diploma.ffzg.hr/ where an independent verification is also possible without scanning the QR code.

2. AI or automation? …

Scanning QR code from the diploma back …

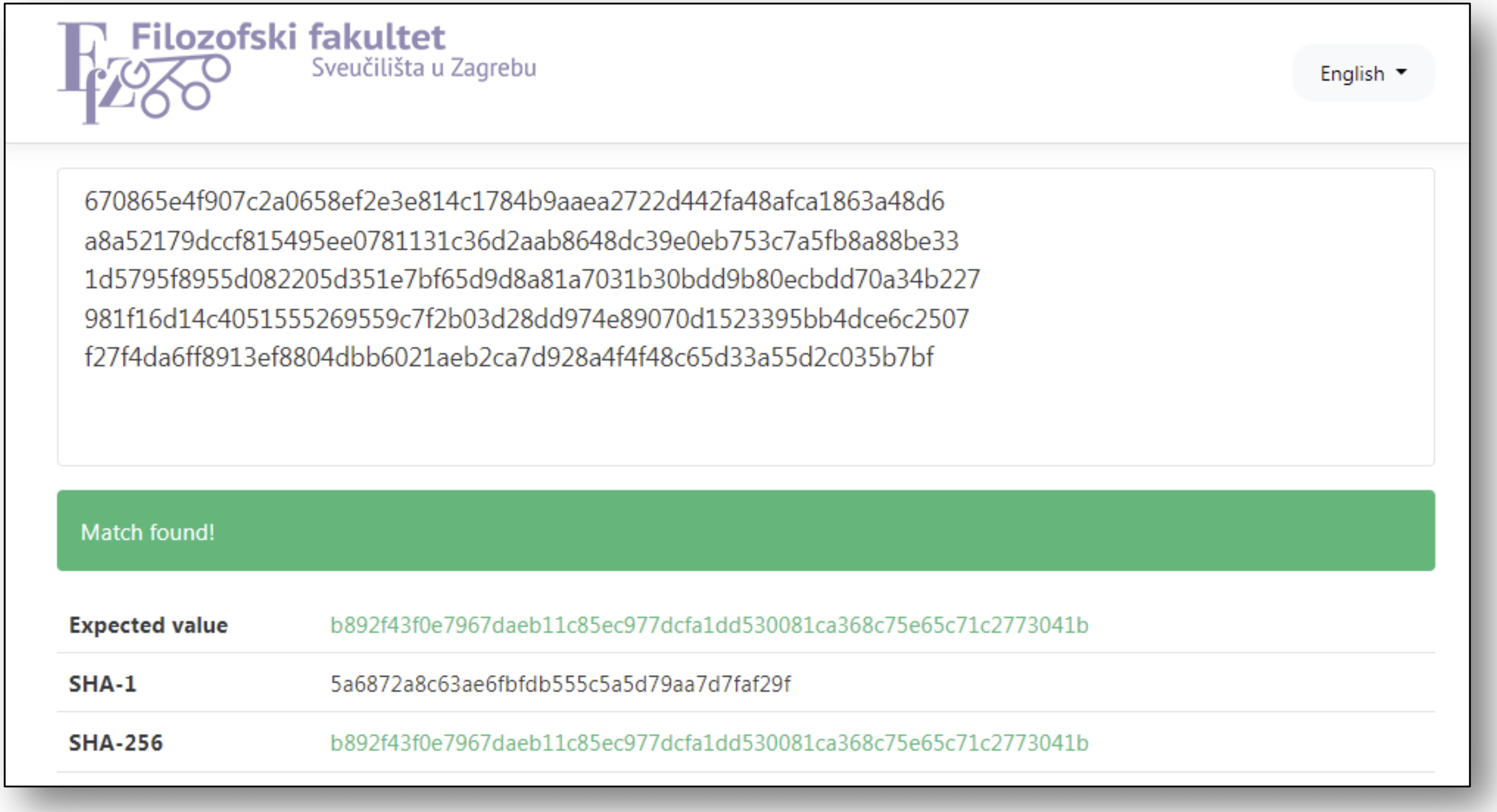


Scanning the QR code allows authentication of a diploma issued after 13 December 2021. The authenticity of the diploma is verified by checking its registration in the blockchain distributed ledger system. Authentication result is displayed on the website: https://provjera-vjerodostojnosti-diploma.ffzg.hr/ where an independent verification is also possible without scanning the QR code.

https://provjera-vjerodostojnosti-diploma.ffzg.hr/

2. AI or automation? …

- Proof from the blockchain distributed system



```
670865e4f907c2a0658ef2e3e814c1784b9aaea2722d442fa48afca1863a48d6
a8a52179dccf815495ee0781131c36d2aab8648dc39e0eb753c7a5fb8a88be33
1d5795f8955d082205d351e7bf65d9d8a81a7031b30bdd9b80ecbdd70a34b227
981f16d14c4051555269559c7f2b03d28dd974e89070d1523395bb4dce6c2507
f27f4da6ff8913ef8804dbb6021aeb2ca7d928a4f4f48c65d33a55d2c035b7bf
```

Match found!

Expected value	b892f43f0e7967daeb11c85ec977dcfa1dd530081ca368c75e65c71c2773041b
SHA-1	5a6872a8c63ae6fbfdb555c5a5d79aa7d7faf29f
SHA-256	b892f43f0e7967daeb11c85ec977dcfa1dd530081ca368c75e65c71c2773041b

3. Research

- InterPARES Trust AI research project's study
  - Identification of critical archival challenges which are the best candidates for improvement by AI technologies in the context of retention and preservation of digital records

3. Research

- Hrvoje **Stancic**, lead & Arian **Rajh + GAAs:** Zeljko **Trbusic**, Vladimir **Bralic**, Patricija **Gligora**, Faculty of Humanities and Social Sciences (FHSS), Croatia
- Alicia **Barnard**, Universidad Nacional Autónoma de México - ENES-Morelia
- Gabriele **Bezzi**, Regione Emilia-Romagna, Italy
- Meltem **Dişli**, Hacettepe University, Turkey
- Pat **Franks**, San Jose State University - School of Information
- Arien **Gonzales Crespo**, El Colegio de México
- Claudia **Lacombe Rocha**, National Archives of Brazil
- Lungile **Luthuli-Ngidi**, University of South Africa
- Patricia (Pat) **Moore**, Carleton University, Canada
- Samir **Musa**, European University Institute - Historical Archives of the European Union, Italy
- Rosely **Rondinelli**, Institute of Technology and Society, Brazil

4. Research – aim and methodology

- Identification of critical archival challenges in the context of retention and preservation of digital records
- Identification of archival challenges arising from digital preservation risks
- Specific factors within challenges will be identified and mapped
- Proposal of how to address them by AI

4. Research – aim and methodology

- Online survey
  - targeted archival practitioners and experts in the field

- Follow up in-person interviews (in progress)

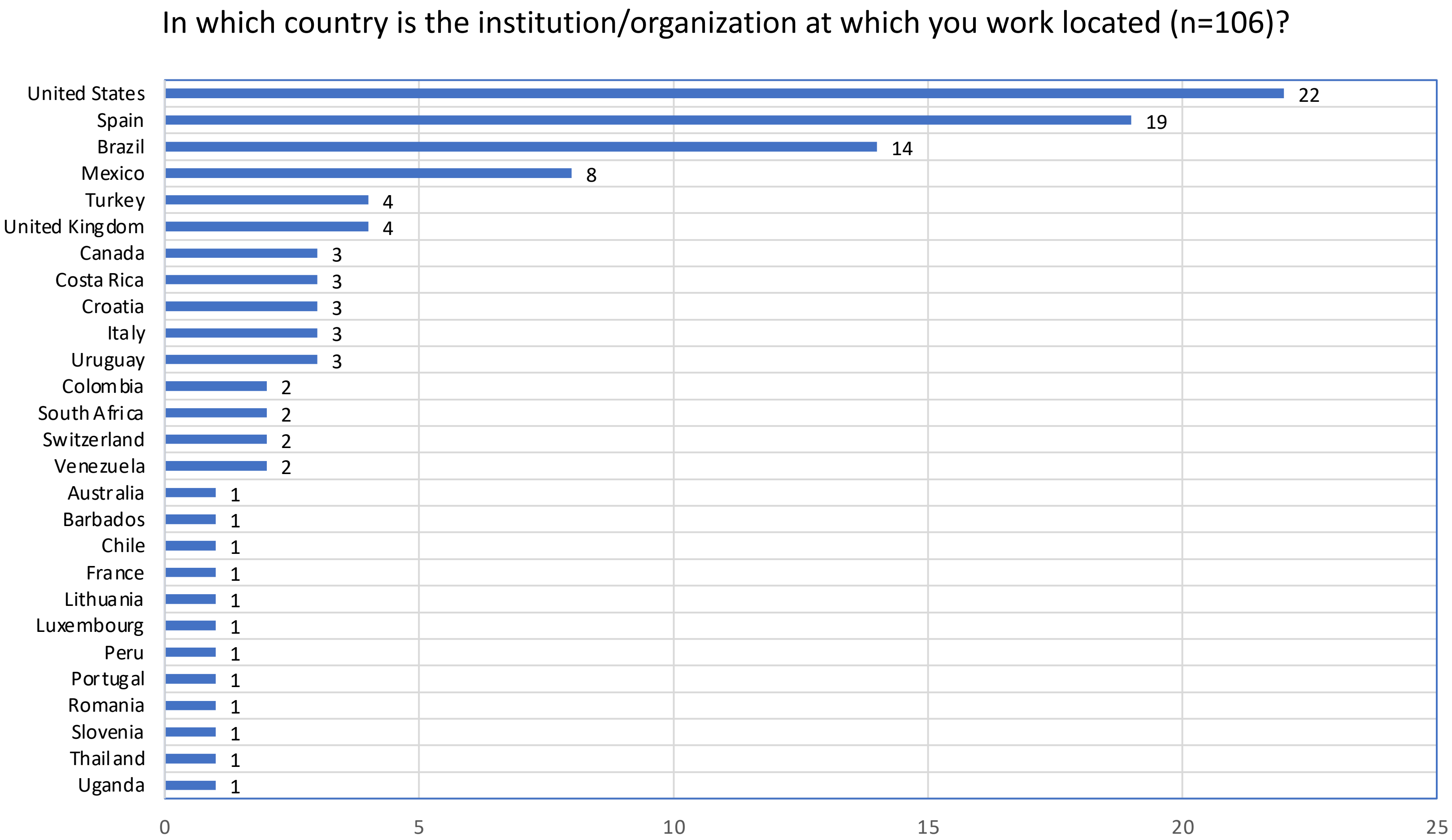
5. Research results

- Online survey
  - 5 March – 9 April 2022 (5 weeks)
  - in English, Spanish, and Portuguese
  - JotForm (https://www.jotform.com/)
- Survey structure
  - 3 parts
  - 14 questions + possible sub questions
- Responses
  - n=106

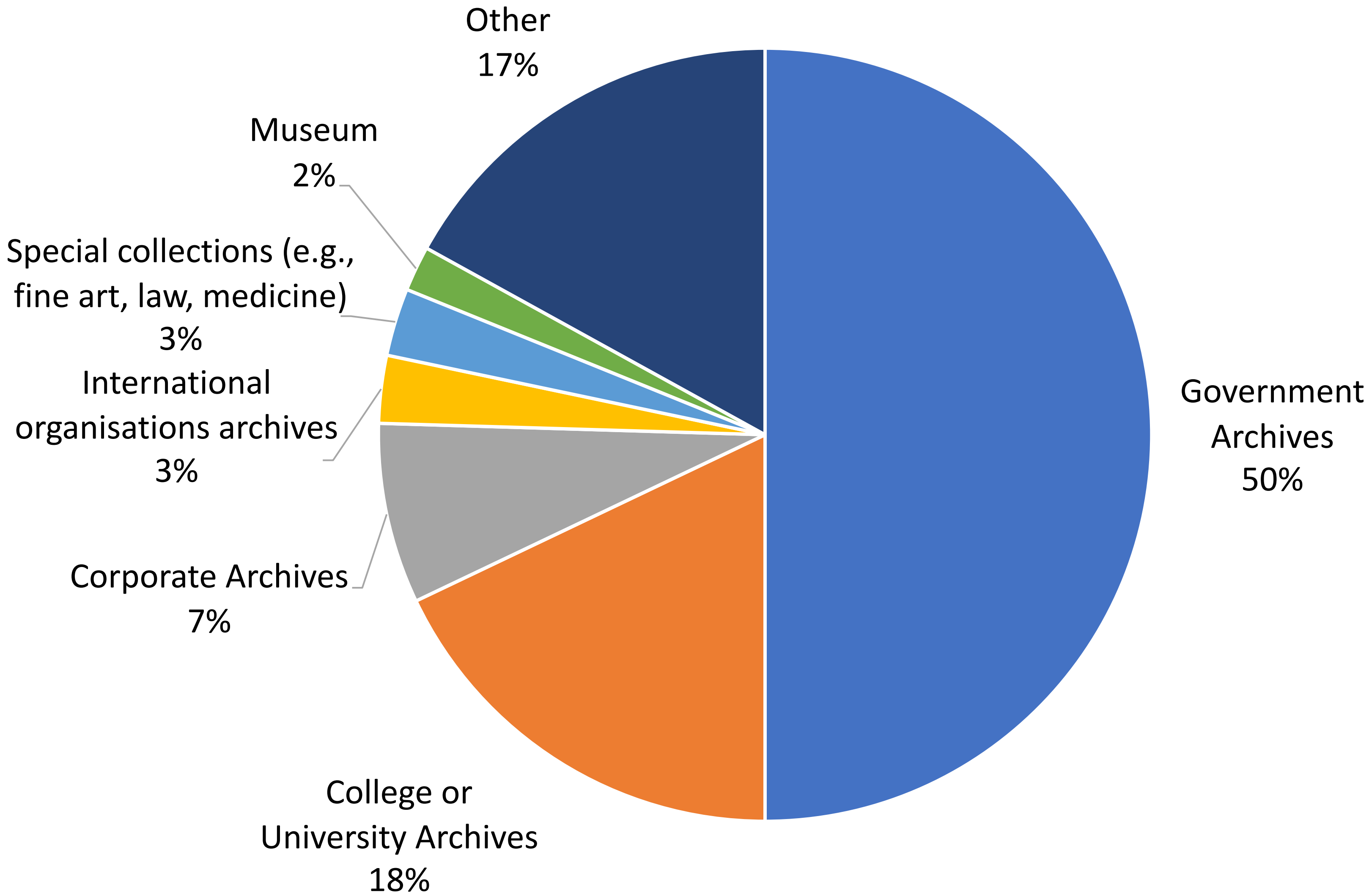
5. Research results ...

- Translation of responses
  - ES → ENG
  - POR → ENG
- Analysis of the results in English

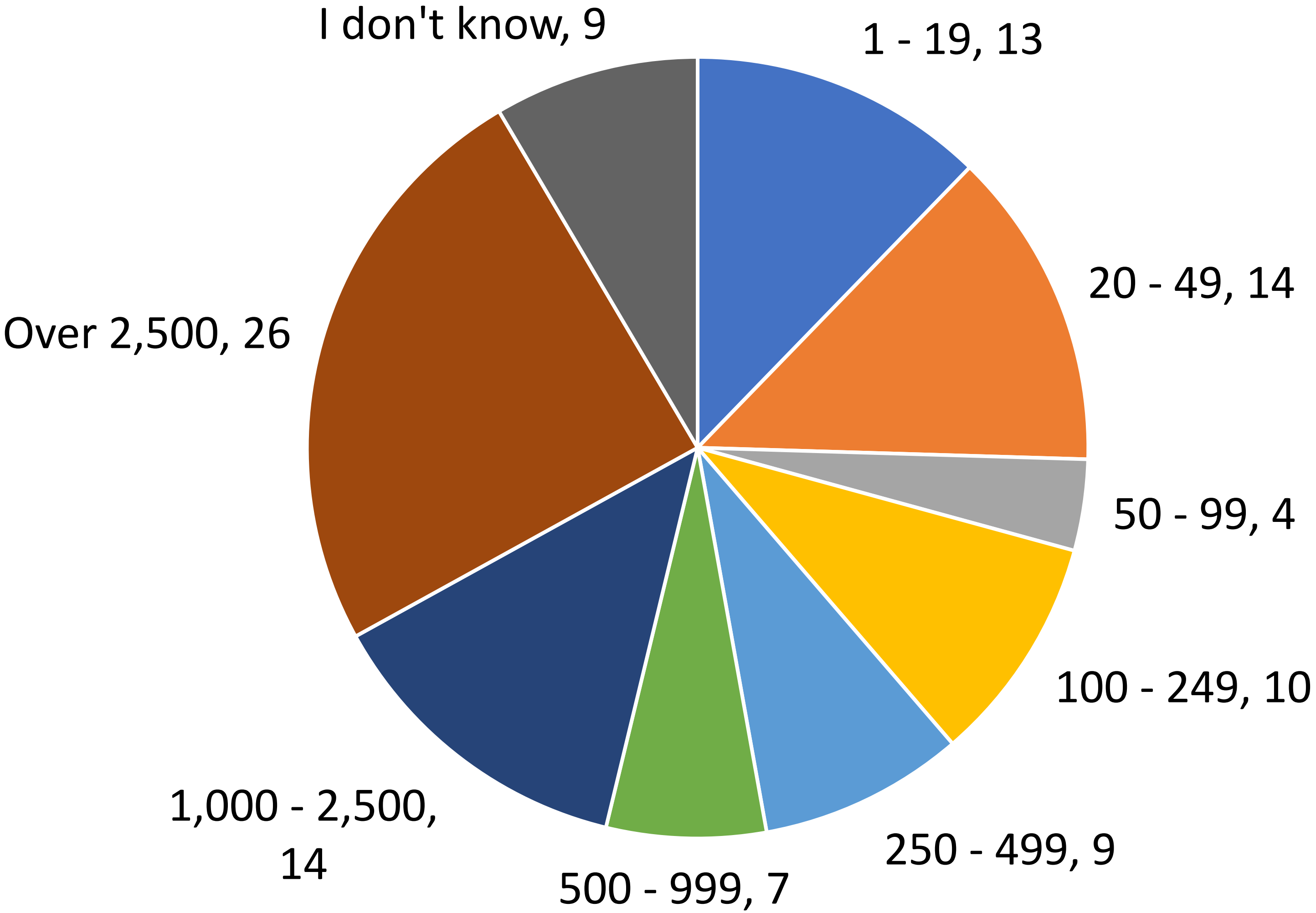
Countries (27)



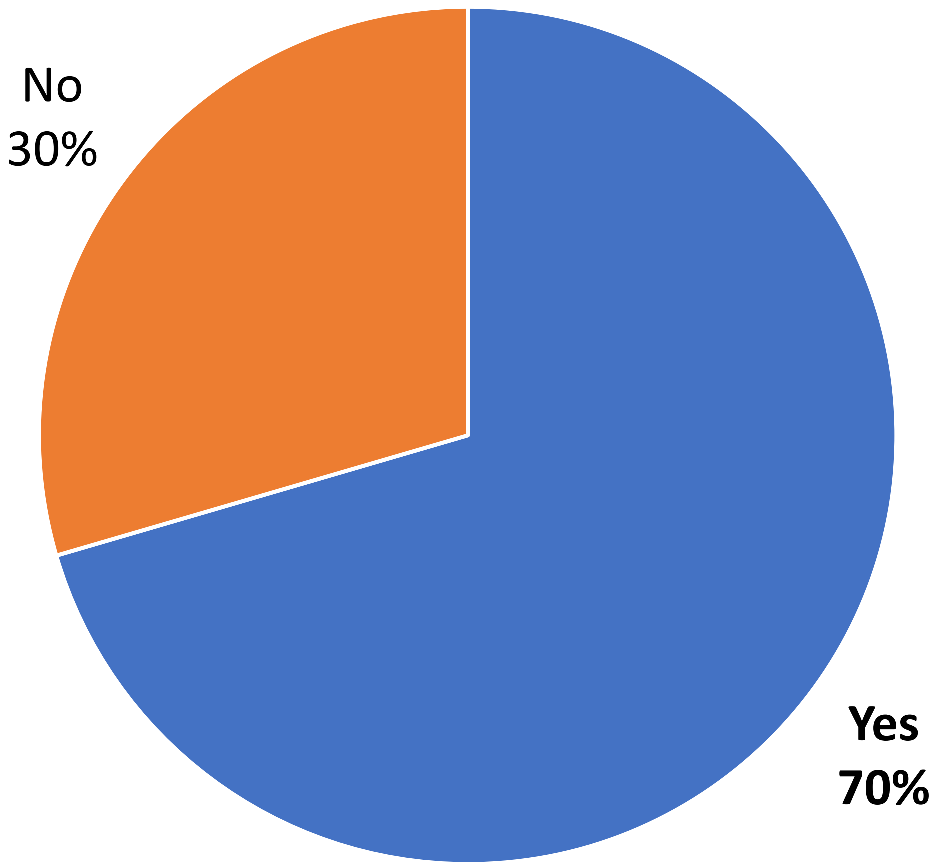
In which type of institution/organization do you work (n=106)?



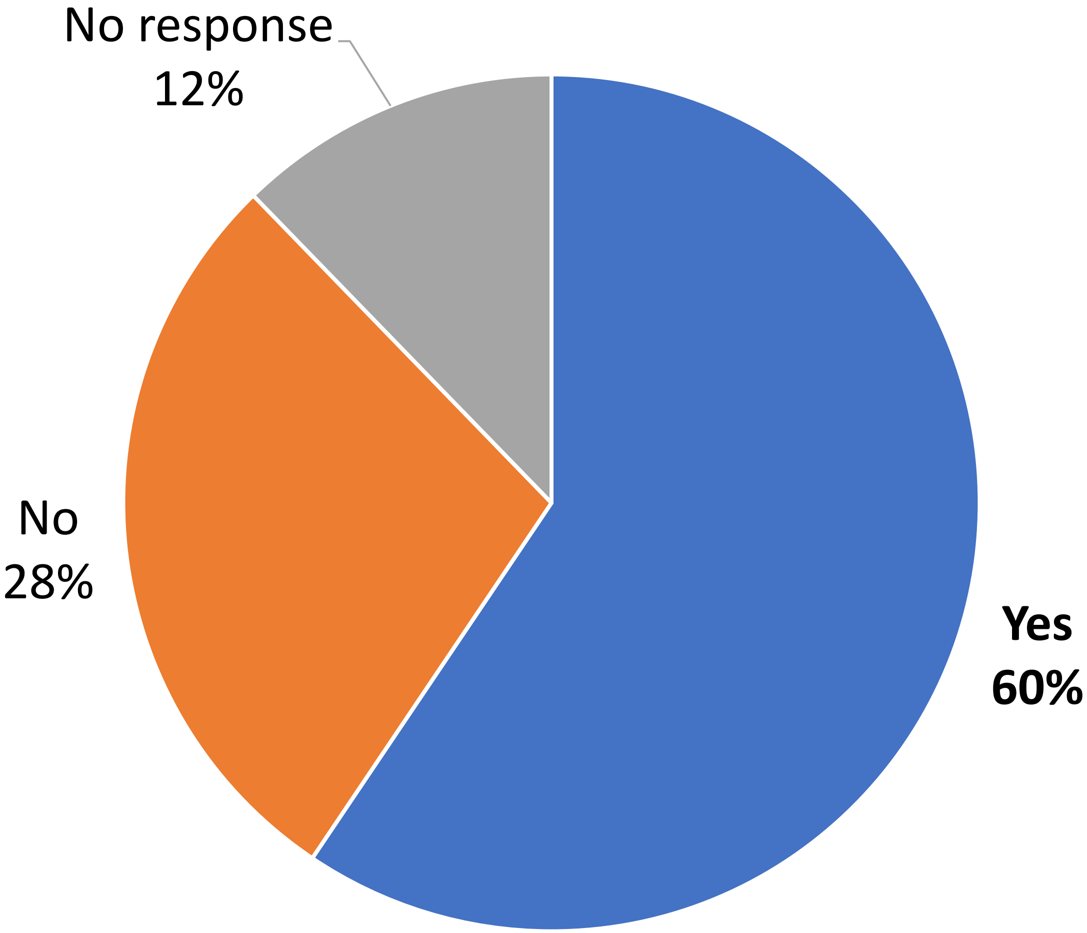
Approximate total number of employees at all locations within the country (n=106)?



Do you perform digital preservation tasks in your institution/organization (n=106)?



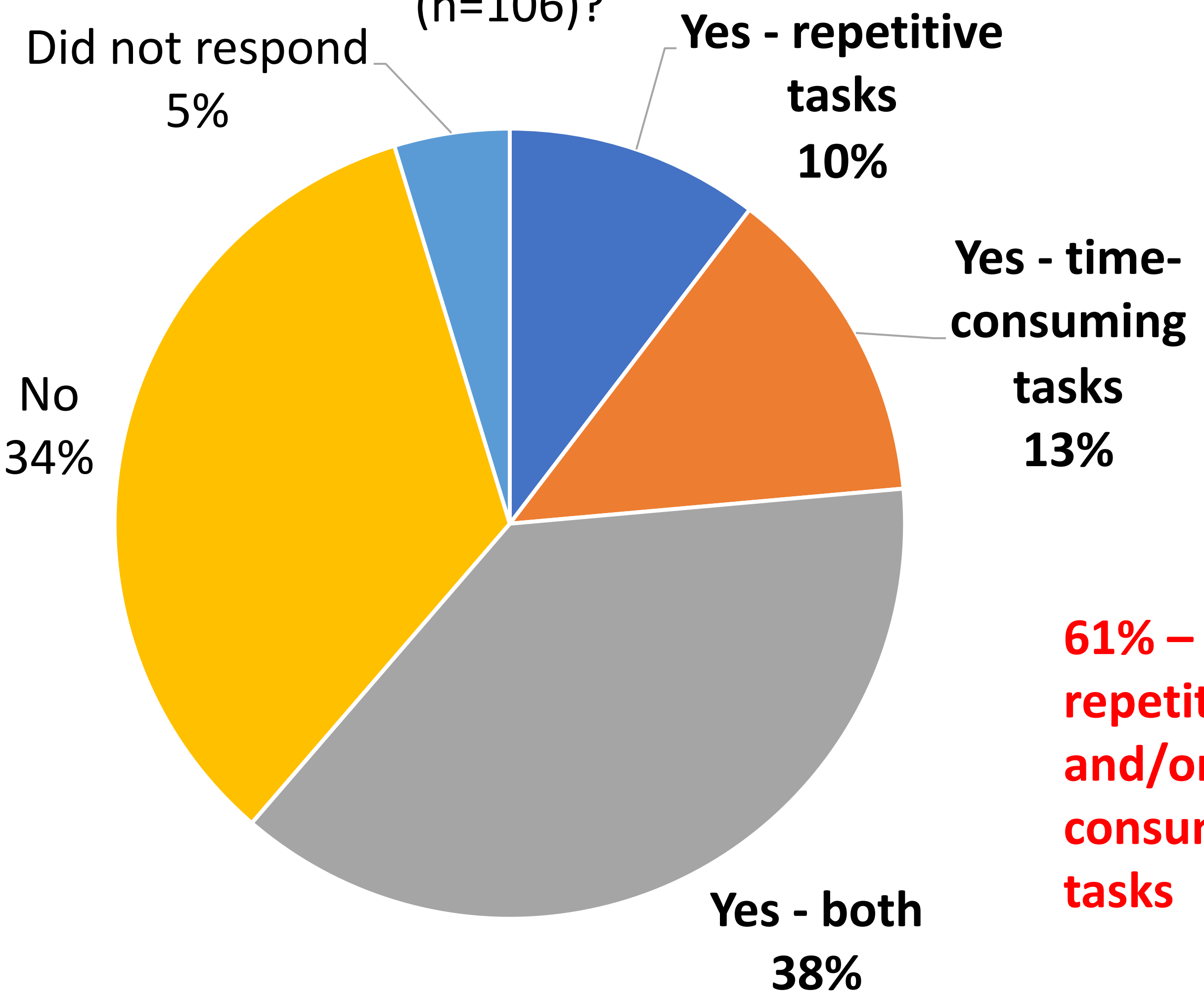
Do any of the digital preservation processes involve large quantities of digital records (n=106)?



What does "large quantity" mean to you (e.g. measured in [giga/tera/peta]bytes, or in number of files)? Please specify, and if possibly elaborate.

- From "numerous files" to 9 PB
  - 9 GB
  - Over 2 billion files
  - A few collections have hundreds of compact discs, or a couple hard drives
  - PB, our holdings are in TB, largest single deposit 130 GB

Do any of the digital preservation processes in your institution/organization involve repetitive or time-consuming tasks (n=106)?

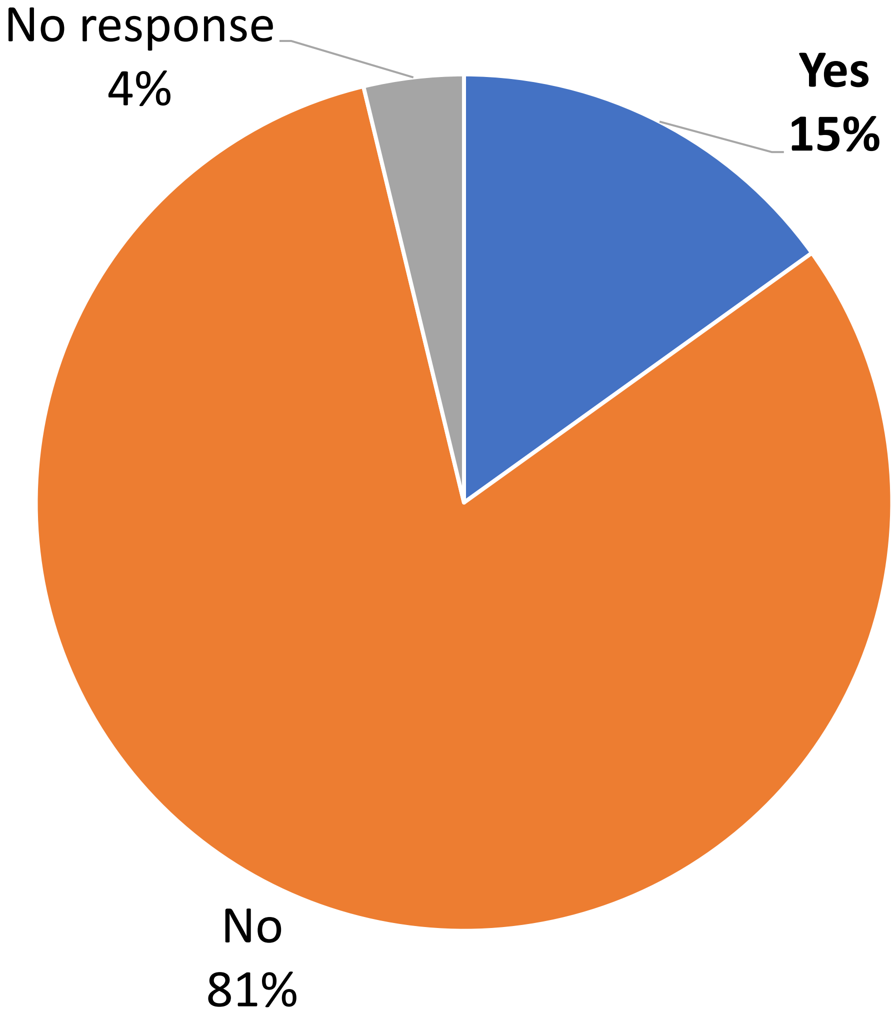


61% – repetitive and/or time consuming tasks

Identified repetitive and/or time-consuming tasks (30 in total, showing 3+)

Adding, gathering, extracting metadata	11
Digitization	10
Capture / ingest	7
File integrity check	6
Indexing	5
Records management	5
Appraisal	4
Backup	3
Renaming files (based on their content)	3

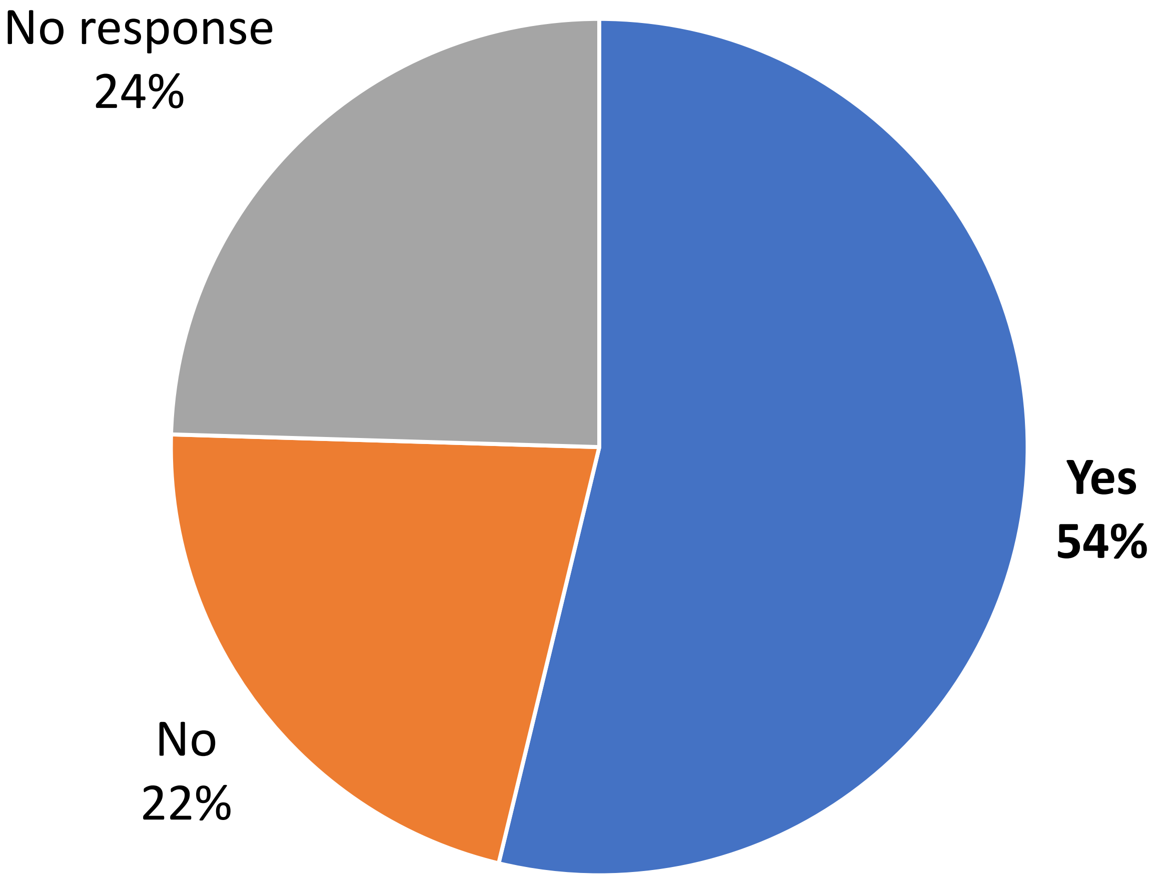
Does your institution/organization use any automated or AI-supported activities in the digital preservation processes (n=106)?



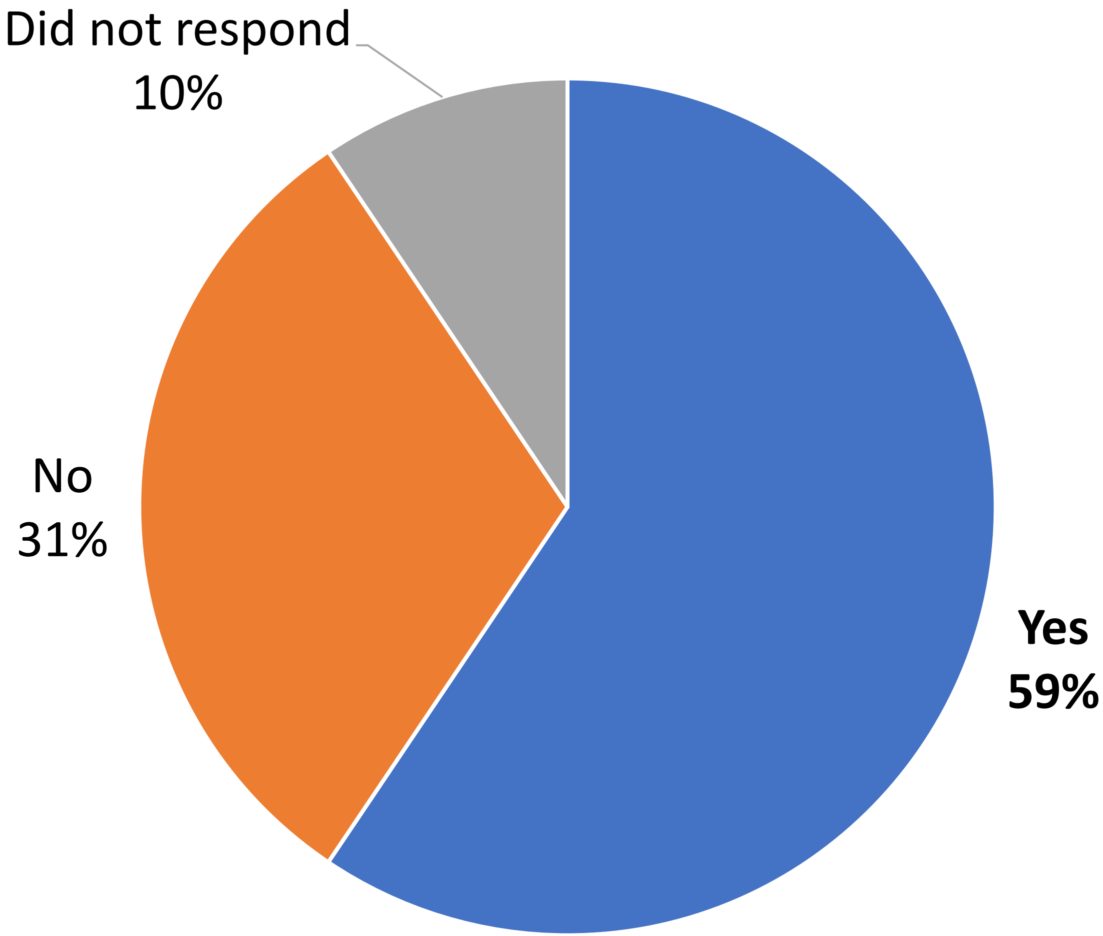
Identified automated or AI-supported activities in the digital preservation processes

Ingest / upload / capture / packiging	3
Classification (and granting access based on it)	2
Metadata operations (description / extraction from records)	2
Search / recommendation engines	2
Software built-in tools	2
Analysis of metadata for PII detection	1
Basic MS Word operations (adding date, grammar check)	1
Data profiling	1
Digitization	1
Format validation	1
ML for identifying born-digital moving image records for preservation	1
Scripts for process automation	1

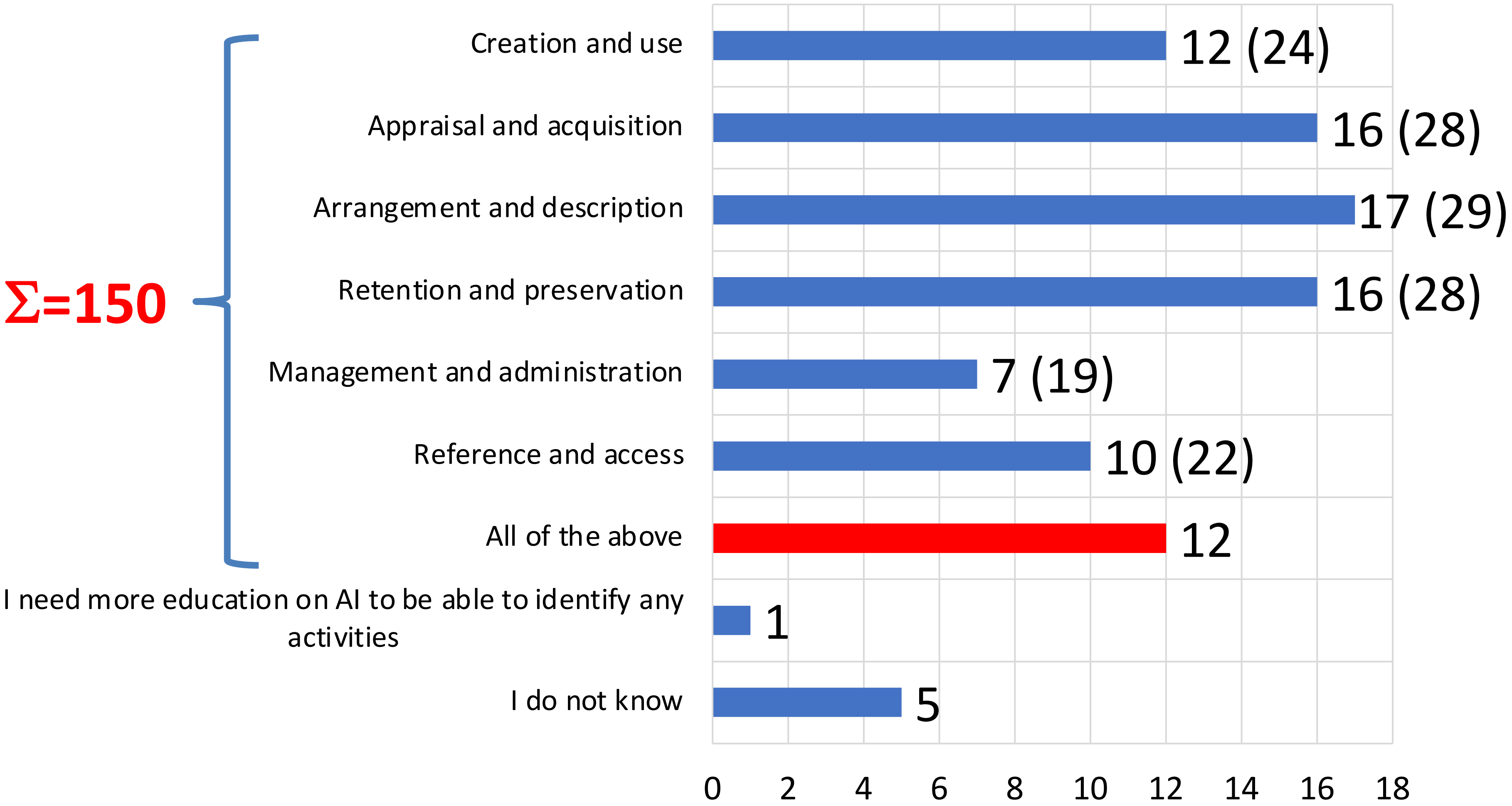
Could the AI-related technologies be integrated into the digital preservation system you are using (n=106)?



Do you have any particular processes which can be (additionally) improved by AI-related technologies (n=106)?



To which group of activities the identified processes which can be (additionally) improved by AI-related technologies best relate to (n=63)?



6. Conclusion

- Critical challenges to be improved by AI
  - Digitisation supported by AI
    - automatic text recognition
    - quality control
    - indexing / classification / metadata extraction
  - Digital preservation supported by AI
    - experience gained from past situations (AI training set) with similar file formats and flagging potential problems with currently kept file formats
  - Reference and access supported by AI
    - identification of PII
    - context-based redaction

THANK YOU!

Digital Preservation and AI - Critical Challenges

Dr. **Hrvoje Stančić**, full prof.
Vice dean for organization and development /
Chair of archival and documentation sciences
Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences
University of Zagreb, Croatia
hstancic@ffzg.hr

LinkedIn