InterPARES Trust AI

AN INVESTIGATION ON THE USE OF AI TO:

- IDENTIFY OR RECREATE ARCHIVAL AGGREGATIONS OF DIGITAL RECORDS
- ENRICH METADATA SCHEMAS

TEAM CU05

Massimiliano Grandi

The TEAM CU05 Study main research question

Can we use **AI tools** to **build** or **recreate archival aggregations** and **metadata** schemas for them?

In environments, **documents are neither classified nor aggregated.** In other cases, aggregations of documents are **not well created**, resulting in an uncontrolled number of documents that are **not sorted**, not placed in the correct folder and **difficult to find**.

In many cases **metadata -** necessary to ensure the reliability, trustworthiness, quality and sustainability of appraisal and acquisition - **are missing**. These problems are particularly serious with regard to **email management**.

Despite progress on various technologies to support document management, software support for those activities remains limited.

Which **AI technologies** could be useful for this purpose for the automatic or semi-automatic management of documents, for example:

- for automatic classification?
- for aggregating the records?
- for filtering and aggregating emails?
- for integrating metadata for describing the creation context and use?
- for automatic appraisal and disposal?

Identification of AI companies

Identification of an initial group of **300 companies** of interest to the study: that group was neither exhaustive nor definitive, but a **starting point**.

Tools for building the list:

- **direct Internet searches** using keywords and text strings;
- **resources and knowledge made available by professionals**
  (Alan Pelz-Sharpe, Andrew Warland, James Lappin, Jenny Bunn and Paul Young)

The group was later limited to **100 companies** on the basis of:

- **statements** where the AI company declares interest for **document management;**
- **expressions of interest** for any aspect of **archives** and **records management**

Finally, from the initial list Team CU05 selected a list of **28 companies** on the basis of:

- their **portfolio**
- their direct involvement in the **record field**
- their **compliance** with regulatory frameworks and standards relevant to the domain
- the general **reputation** of the company.

Companies that replied to the survey

13 companies replied:

Collabware (Canada)

Iron Mountain (USA)
Bis (USA)

Aluma (UK)

Cortical.io (Austria)
Read-Coop (Austria)

Anzyz Technologies AS (Norway)

Castelpoint Systems (Australia)
RecordPoint (Australia)

Grupo Adapting (Spain)

expert.ai (Italia)
Quest-it (Italia)

Bizamica (India)

What services are delivered

They may be grouped in **5 categories**:

- **Automatic indexation / classification** - this seems to be by far the most advertised function;
- **Automatic data extraction** - when papers are considered, often at the same time as a document is being scanned;
- **Intelligent processing** - i.e. the application starts and advances processes automatically on the basis of features detected on the document, e.g. route documents to specific people or implements retention schedules;
- **Intelligent discovery** – information retrieval by e.g. comparing documents or analyzing concepts;
- **Automatic redaction** (relating to data protection)

Questionnaire and interviews

In order to gather more precise information, we prepared a **questionnaire containing 25 questions** divided into four sections and aimed at collecting systematically the information for an adequate assessment of the applications

The questionnaire was explained orally during a **preliminary meeting** with information management staff and software engineers.

Subsequently, the companies filled out the questionnaire available on **Google Forms**

I SECTION	achievements
II SECTION	specific capabilities (for recordkeeping and email systems)
III SECTION	technologies and methods used in the AI applications
IV SECTION	audit-checks - key performance indicators

Survey outcome from the archival perspective (1/4)

Automatic classification (yes: 10 – no: 3)

- analysis of **metadata elements** available both in the records and aggregations (case-folder specifications)
- identification of **document type**
- in case the available metadata should prove to be insufficient for classification, then classification is based on the **record content**
- generation of **labels and tags** belonging to any record classification scheme (taxonomy or term ontology)

Filing / Aggregation (yes: 10 – no: 3)

- by **document type**
- by considering the original structure of the content source
- generation of **labels and tags** from any record classification scheme, based on the **record content**

Inferences on records grouping (yes: 9 – no: 4)

- based on **content and/or context**
- if there is **metadata** to represent those processes (e.g. a case file number)

Inference on organisation or person (yes: 7 – no: 6)

- If the involved entities are stated in the **content** of the document

Re-establish the archival bond of unarranged records (yes: 5 – no: 8)

- extraction of data **from records content or from metadata**

Indexing to create links, aggregations (yes: 6 – no: 7)

- difficult task **if contextual data is lost**

Survey outcome from the archival perspective (2/4)

- The role of any **metadata fields found or inferred** is always at the center of any reply.
- The **records typology** – when available – is often considered another crucial component for the successful application of the AI techniques to the records.
- In terms of **records classification**, only one company said its platform **can be trained by the users thanks to a specific set of data** for generating autonomously labels and tags related to any record classification scheme understood as based on taxonomy or term ontology.
- In the other cases **the human intermediation is considered not replaceable** for providing consistent results.
- In terms of **records aggregation** or **re-aggregation**, the promises for automatization are not very encouraging, as this possibility is confirmed to be **limited to very specific cases** such as
  - ✔ **defining records types**, when the **users' specifications are already in place**, or
  - ✔ **establishing functional relations among records** when the **original structure** of the content source already provides **basic *intelligent* information**.
- The automatic or semi-automatic aggregation based on the document content is **only suggested** and is usually **supported by user validation**, of **human-in-the-loop workflow** or **rules available at the creation**

Survey outcome from the archival perspective (3/4)

As to **provenance, it seems not to be easily recognizable** by AI solutions **when based on inferences and without very specific requirements** such as

- the identification of the right case-folder,
- the presence of a stamp, a statement clearly expressed in the record,
- specific metadata and/or classification elements.

Also the **reconstitution of the archival bond** – when lost or not explicitly defined – is recognized as a complex activity, without the significant help of users and/or consistent descriptive information available and, in any case, it implies **more investments, not yet supported by the market**

Survey outcome from the archival perspective (4/4)

- What has been found out testifies that the complexity of archival functions cannot be easily reduced and removed by an automatic approach, but only supported by the AI technologies through the **intermediation by users and professionals**.
- The **terminology is a crucial challenge**.
- As a consequence, when interacting with market players involved in the implementation of AI platforms, the archival community must pay a lot of attention
  - to clarify their concepts behind general terms such **aggregation** and **classification** and
  - to correctly interpret AI expression such, "**Intelligent Document Processing**" which, usually, has nothing to do with document, with its processing and, at the end, with ***archival intelligence*.**
- Our research is only at its first phase but we have already recognized that we can and must accept the challenges without being intimidated:
  - by **the pressure of top management** asking for archival miracles based on new disruptives technologies;
  - by **AI market players promises** which have to be carefullly checked;
  - by the **complexity of AI technologies**, because the solutions they offer imply more than in the past our knowledge and experience

Case Study 1 – NATO (1/7)

- NATO Archives was established in 1999 and is based in Bruxelles; its main scope is the public disclosure of Alliance records older than 30 years. It mostly keeps documents of the North Atlantic Council and its sub-committees, the Military Committee and its working groups. Committees are created to address specific topics before reporting back to the North Atlantic Council.

- The fonds/series structure reflects the committee structure and it is arranged based on the reporting chain within the organization and chronologically. Records are referenced by an alphanumeric string as per the following slides

- NATO Archives has made available for the case studies a large series of declassified Committee records from the Fifties to the Nineties. The Committee records are all scanned and OCRed in PDF formats.

- The company providing the technology for the case study is RecordPoint, from Australia. RecordPoint made available for the study its platform called Records365. Records365 is an in-place records management platform that can ingest data from any digital content source

Case Study 1 – NATO (2/7)

- Records365 classifies records to determine their lifetime and will dispose them when disposal is due. Automated classification follows two approaches: 1) expert system (rules-based) classification that uses record metadata, and 2) machine learning classification, based on record text. In Records365 Machine learning processes use a supervised learning strategy.
- Records365 can also use machine learning to categorize records in a way that matches a given classification taxonomy, and is able to enrich records by extracting PI, named entities, and other signals from text and metadata content.

Case-study proposed deliverables

2. *The application is able to aggregate the digitised documents according to clusters, such as for example series/creators/topics-objects/identifier/signatories …*
3. *The application is able to capture additional metadata, i.e. signatories, the original security classification of the document, the public disclosure notice, agenda items*
4. The application performs text summarization on selected items and/or series of documents
5. The application can flag items that are not NATO documents (e.g., national documents)
6. The application proposes the semantic tagging of the given collection according to controlled vocabulary/ontologies including events/places/people

Case Study 1 – NATO (3/7)

Deliverable #1 (*The application is able to aggregate the digitised documents according to clusters, such as for example series/creators/topics-objects/identifier/signatories*) - some contextual information to assess it:

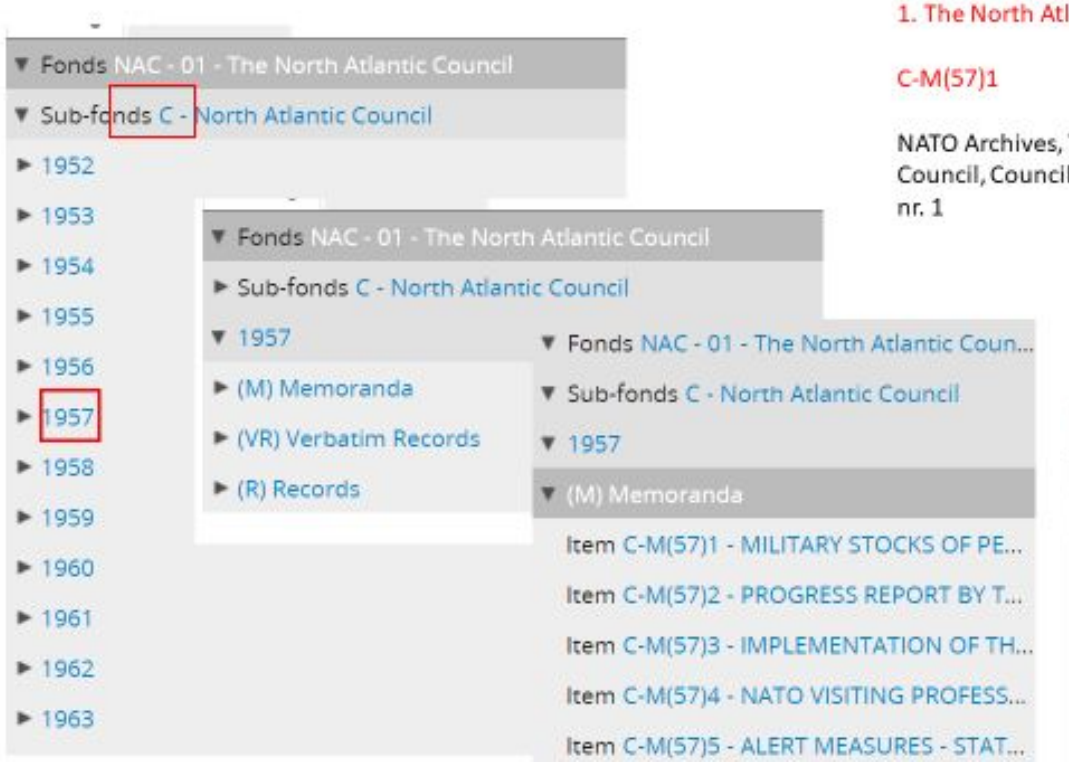
- A minimum of 50 sample documents classified by the relevant ontologies to be able to perform automatic classification using AI/ML. This is necessary to building a machine learning model in the system
- The series to train the model are chosen based on how many records were available in each of them
- The top 18 categories were chosen based on how recent they are and on file size
- Ca 14000 records ingested into the platform so far
- Access to the platform to review the results has been provided to InterPARES member

Case Study 1 – NATO (4/7)

Some observations on Deliverable #1:

- Providing traditional records management toolkits such as ontologies, filing plans, taxonomies is a pre-requisite. Limitations on the archives side, as some materials not disclosed to the public.
- 14000 records used so far out of a potential number of 300000. The creation of zipped packages of records is a very manual and lengthy process that archivists had to face
- The testing of the platform is limited to the chronological series of Committee documents, because in this case the archives could provide 50 instances of each kind of record
- NATO also holds curated collections on specific themes/topics. These collections are made of records from different series and originators. It hasn't been possible to start with the testing on these files, because there were not enough instances of the same kind of record to train the algorithm, so the minimum requirements were not met
- RecordPoint currently supports English only for its AI/ML Classification. Half of NATO records' corpus is in French. It worked quite fine also for French records
- Metadata enrichment. So far PII (Personal identifiers information) has been proposed by the platform, but it resulted in many false positives. Review is needed

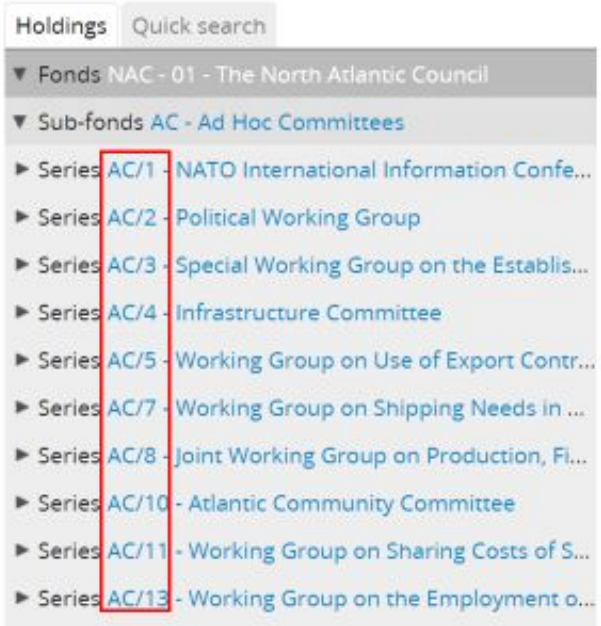
Case Study 1 – NATO (5/7)



1. The North Atlantic Council Series

C-M(57)1

NATO Archives, The North Atlantic Council, Council Memoranda, 1957, nr. 1

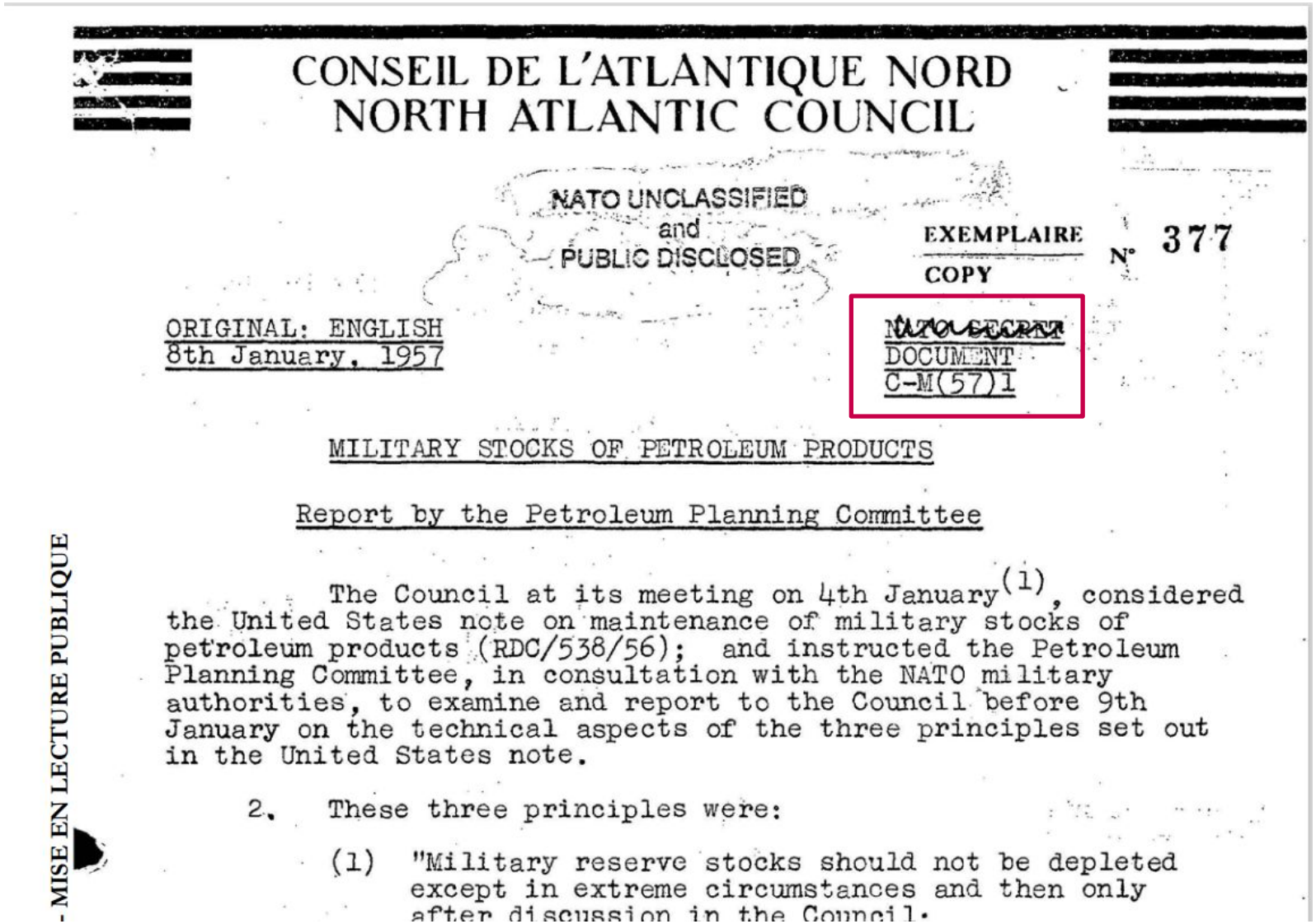


2. The Subordinate Commitees series

AC/4-WP-349

NATO Archives, Infrastructure Committee, Working Paper,

Case Study 1 – NATO (6/7)



CONSEIL DE L'ATLANTIQUE NORD
NORTH ATLANTIC COUNCIL

NATO UNCLASSIFIED
and
PUBLIC DISCLOSED

EXEMPLAIRE / COPY N° 377

ORIGINAL: ENGLISH
8th January, 1957

~~NATO SECRET~~
DOCUMENT
C-M(57)1

MILITARY STOCKS OF PETROLEUM PRODUCTS

Report by the Petroleum Planning Committee

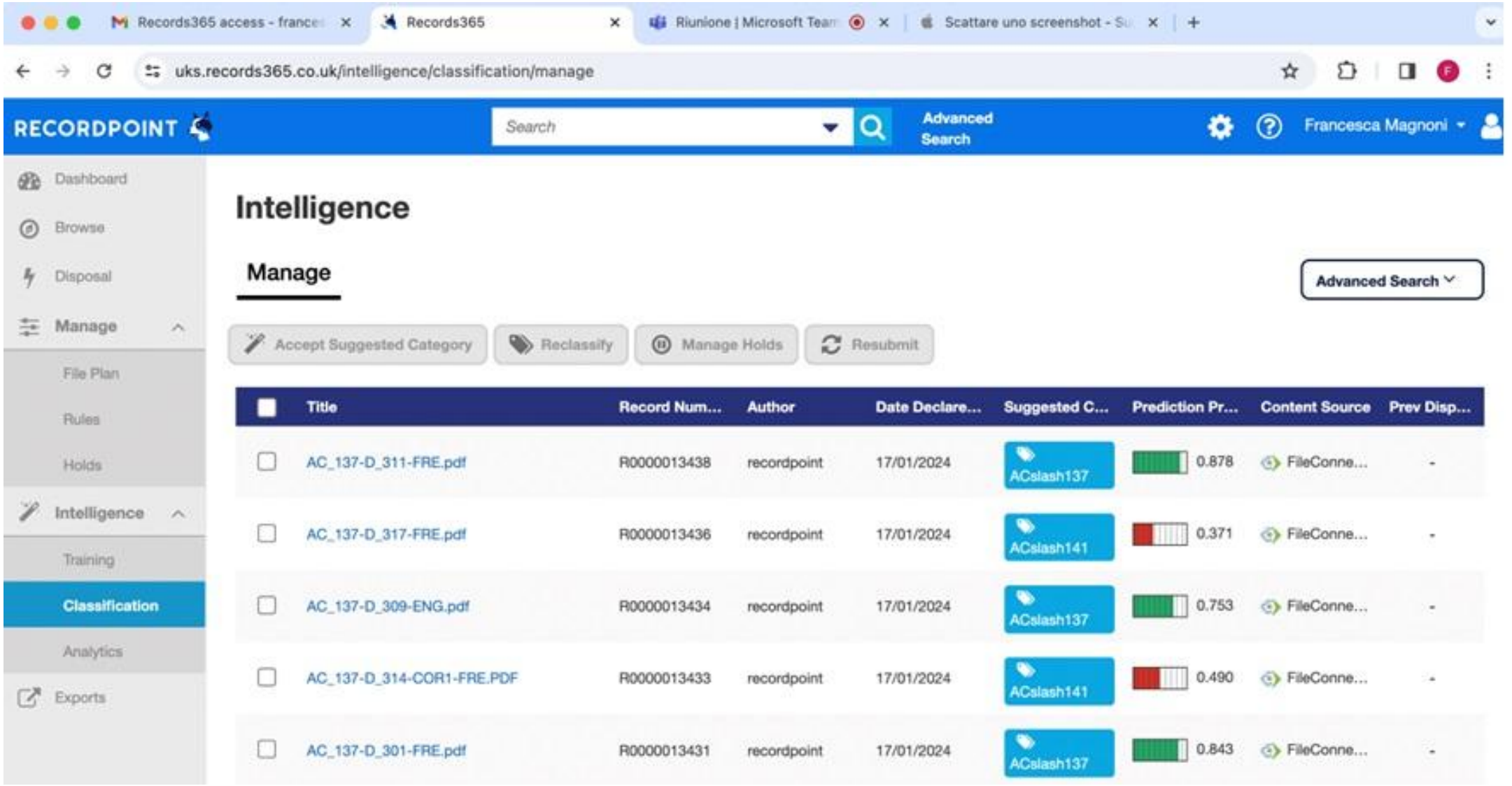
The Council at its meeting on 4th January(1), considered the United States note on maintenance of military stocks of petroleum products (RDC/538/56); and instructed the Petroleum Planning Committee, in consultation with the NATO military authorities, to examine and report to the Council before 9th January on the technical aspects of the three principles set out in the United States note.

2. These three principles were:

(1) "Military reserve stocks should not be depleted except in extreme circumstances and then only after discussion in the Council.

- MISE EN LECTURE PUBLIQUE

Case Study 1 – NATO (7/7)



Case Study 2 – CISU (1/4)

- CISU stands for Centro Italiano Studi Ufologici (Italian Center for UFO Studies).
- Largest archives concerning UFO sightings and UFO studies in Italy; second largest one in Europe (after AFU - Archives for the Unexplained - in Sweden).
- CISU has made available for the case studies its large series of news clippings from the Fifties of the 20th Century to date (ca. 70,000 – 80,000 documents). The news clippings are scanned images of paper news clippings (mostly .PDF and .TIFF formats).
- The company providing the technology for the case study is Anzyz, from Grimstad, Norway.

AIR FORCE PLANE HUNTS FOR UFOs

- 3 JAN 1979

A Royal New Zealand Air Force Orion anti-submarine reconnaissance aircraft was patrolling the night sky above Kaikoura early today in search of unidentified flying objects.

The Orion is equipped with radar, cameras and a host of other sophisticated tracking and monitoring equipment.

Its mission was to patrol the north-eastern region of the South Island from about midnight till just before dawn in an official attempt to solve a series of mysterious sightings during the last 12 days.

The crew, under Squadron Leader R. J. Carran, will be in constant touch with civil aviation radar operators at Wellington who claim to have tracked scores of unidentified flying objects over the Clarence River area.

Air Force operations decided yesterday morning to put the aircraft into the air as the debate grew over UFO sightings made by Safe-Air pilots on December 21 and again early last Sunday morning.

No Threat

A Ministry of Defence spokesman, Squadron Leader G. T. Clarke, said last night the sightings were not regarded as a threat to the country but they were interesting enough to warrant investigation.

peared to be near Wellington Airport, it was believed to have been one set off accidentally.

But defence public relations officer, Squadron Leader G. T. Clarke, said the search and rescue organisation was being alerted.

Worldwide attention was focussed on the sky above Kaikoura after a film of Sunday morning's sighting beamed out over British, American and Australian television on Monday night.

The film was shot from a Safe-Air Argosy freighter by New Zealand cameramen accompanying a journalist working for a Melbourne television channel. New Zealand viewers saw the film last night.

An international debate has erupted over possible explanations of the large, white, glowing ball captured on film.

And it has sparked off a flood of reports of sightings of UFOs in Britain and of balls of red light and cigar-shaped discs spitting fire over Italy.

A British astronomer, Sir Bernard Lovell has dismissed the objects as "most

The image filmed by the Melbourne television channel over New Zealand's east coast.

appear to be moving at about the same speed because of its great distance.

Pilot Sure

The pilot of the Argosy from which the television film was shot and the first

Venus is at present crescent-shaped and these were round."

But, according to the Auckland Observatory last night, Venus remains a bright light to the naked eye and the crescent shape

Society of New Zealand, Mr D. Calder, said yesterday that the objects were "clearly not astronomical."

He said many UFO reports could be traced to meteors, bright planets and bright stars seen in unusual circumstances, but this was not the case with the recent sightings.

"All I can say is that they were not astronomical objects, but the evidence does not preclude other natural phenomena."

Adding to the mystery, the director of the physics and engineering laboratory of the Department of Scientific and Industrial Research, Mr M. A. Collins, said that it would be stretching the imagination to think of any natural causes that could explain the sightings.

He said the fact that the objects registered on radar meant that they were "electro-magnetically solid." The least they could be was a dense mass of electro-magnetic particles.

Difficult

And a Canterbury University scientist, Mr J. A. Webb, who has been making a long study of UFOs, said it was

Case Study 2 – CISU (2/4)

- Anzyz has made available its platform Corpus Cube Linguistics (CCL™), based on unsupervised learning.
- To carry out the case study, Anzyz has initially parsed the news clippings by OCR by using another application not related to CCL™ and not based on AI.
- The signature feature of CCL™ is its capability to analyze the unstructured contents of the documents fed to it and then build indexes and proposals of categorisation which then humans can assess and correct by providing the platform with relevant feedback.
- In order to enable Anzyz's platform to show fully its affordances, it is necessary to supply it with a very large body of documents, ideally at least 100,000 documents, as the more numerous the documents fed to the platform are, the more effective the platform may become (because of the size of the training ground).

Case Study 2 – CISU (3/4)

Some of the first outcomes of the case study:

- Corpus Cube Linguistics (CCL™) can create indexes of elements such as dates and names of people and places
- It identifies names using concepts as well as personal identifiable information elements.
- By its very design, however, CCL™ is not able to leverage on finding aids and lists prepared by humans - it carries out this job by analysing the documents and data and creating conceptual graphs on its own.
- CCL™ follows the same strategy when it groups and categorises documents; therefore, when a researcher from CISU provided Anzyz with a list he himself had created by dividing into 6 categories ca. 3,000 news clippings, Anzyz replied the list was good only as a benchmark for humans, to check the job done by CCL™.

	A	B	C	D	E
1	**DATE**	**PERIODICAL NAME**	**ARTICLE TITLE**	**CLASS**	**Sightings reported**
2	00-01-1978	Selezione dal Reader's Diges	I LIBRI - FUOCO DALLO SPAZIO	STO	YES
3	02/01/1978	Eco di Genova e della Liguria	LA FINE DI MANTELL	STO	YES
4	02/01/1978	Informatore (L')	UNA LETTERA DI PROVERBIO SUGLI UFO	LET	
5	02/01/1978	Stampa Sera	NELLA BIBBIA EXTRATERRESTRI	REC	
6	03/01/1978	Stampa Sera	UNA SERA TUTTA TRANQUILLA	REC	
7	03/01/1978	Alto Adige	"VOCE SPAZIALE" TERRORIZZA TELESPETTATORI INGLESI	NOT	
8	05/01/1978	ANSA	AEREO MILITARE E NON "UFO" SOPRA CAGLIARI	NOT	YES
9	06/01/1978	Corriere della Sera (Il)	ERA UN AEROPLANO L"UFO" AVVISTATO SOPRA CAGLIARI	NOT	YES

Case Study 2 – CISU (4/4)

- CCL™, therefore, is not able to benefit much from already available finding aids and reference information; that might mean it is effective mainly in situations where the reconstruction of the archival bond and the identification of the metadata elements are to start from scratch (when actually an application that can try and guess connections among concepts and contents might come in handy).
- Likewise, because of its very design and purposes, CCL™ seems to be unable to analyse recurring patterns of extrinsic elements of the documentary form: e.g. cannot examine the structure of textual labels - appended some decades ago to the paper news clippings and featuring a recurring sequence of elements - and identify the first and last line of the labels, where pieces of information important for CISU are contained.

LA NAZIONE
VIA FERDINANDO PAOLIERI 2
50121 FIRENZE FI
Dir.Resp.GABRIELE CANE'
ARRETRATO PER DISGUIDO POSTALE
Data:26 Maggio 1992

Ed. SARZANA

«EMANAVA UNA STRANA LUCE ED È SPARITO IN POCHI SECONDI IN DIREZIONE DEL MARE»

Vallecchia, Ufo nella notte

La misteriosa apparizione nel racconto di una famiglia

Servizio di
Manuel Cecchinelli

Ufo a Vallecchia? Notizie di questo genere vanno sempre assunte e riferite con tutte le cautele del caso, ma sembra davvero che misteriosi «oggetti volanti non identificati» siano stati avvistati dalla piccola frazione di

apparso e scomparso in un lasso di tempo brevissimo: in tutto non più di venti secondi. Naturalmente non sono pochi coloro che commentano sostenendo che possa anche trattarsi di qualche scherzo di... Bacco. Ma, a parte che la «visione» è stata collettiva e quindi più attendibile, bisogna dire che

Thank you!

Any comments are welcome!