Further with Knowledge Graphs. M. Alam et al. (Eds.)
AKA Verlag and IOS Press, 2021

doi:10.3233/SSW210045

BESOCIAL: A Sustainable Knowledge Graph-Based Workflow for Social Media Archiving

Sven LIEBER [a,1], Dylan VAN ASSCHE [a], Sally CHAMBERS [b,c], Fien MESSENS [c], Friedel GEERAERT [c], Julie M. BIRKHOLZ [b,c] and Anastasia DIMOU [a]

[a] *Ghent University – imec – IDLab, Department of Electronics and Information Systems, Technologiepark-Zwijnaarde 122, 9052 Ghent, Belgium*
[b] *Ghent Centre for Digital Humanities, Ghent University, Ghent, Belgium*
[c] *KBR Royal Library of Belgium, Brussels, Belgium*

Abstract. Social media as infrastructure for public discourse provide valuable information that needs to be preserved. Several tools for social media harvesting exist, but still only fragmented workflows may be formed with different combinations of such tools. On top of that, social media data but also preservation-related metadata standards are heterogeneous, resulting in a costly manual process. In the framework of BESOCIAL at the Royal Library of Belgium (KBR), we develop a sustainable social media archiving workflow that integrates heterogeneous data sources in a Europeana and PREMIS-based data model to describe data preserved by open source tools. This allows data stewardship on a uniform representation and we generate metadata records automatically via queries. In this paper, we present a comparison of social media harvesting tools and our Knowledge Graph-based solution which reuses off-the-shelf open source tools to harvest social media and automatically generate preservation-related metadata records. We validate our solution by generating Encoded Archival Description (EAD) and bibliographic MARC records for preservation of harvested social media collections from Twitter collected at KBR. Other archiving institutions can build upon our solution and customize it to their own social media archiving policies.

Keywords. Social Media, GLAM, Knowledge Graph, RML

1. Introduction

The web, and in particular social platforms, have become social infrastructures for public discourse [1,12] which serve as records of the past. However, these records are usually centrally maintained by profit-based social media providers and, thus, preservation by third parties is necessary.

Data preservation is a resource expensive task which requires long term commitment involving software, data and human resources [3]. Social media poses preservation

[1]Corresponding Author: Sven Lieber, Ghent University – imec – IDLab, Department of Electronics and Information Systems, Technologiepark-Zwijnaarde 122, 9052 Ghent, Belgium; E-mail: Sven.Lieber@ugent.be

challenges: non-technical experts of the GLAM domain[2] have to select harvesting tools, and social media consists of dynamic content[28] and heterogeneous data formats which have to be adequately processed and described.

Furthermore, preservation-related metadata for social media is also heterogeneous, aggravating interoperability and data stewardship. Usually metadata documents which describe collections allow efficiently identifying sources [3]. Yet, different preservation systems may require metadata in different syntax which also represent different perspectives. For example, MARCXML[3] records from the library domain may be used to describe a social media collection from a bibliographic point of view, whereas Encoded Archival Description (EAD) [4] XML records from the archive domain may be used to describe the collection's content hierarchically in more detail. This hampers data stewardship because there is no uniform and interoperable description of the preserved social media collections, let alone provenance of the collection process itself which is crucial[21,28].

Semantic Web and Knowledge Graphs are promising solutions in the GLAM domain [2] as they enable applications across heterogeneous data and address the mentioned issues. However, existing approaches [7,18] assume already curated metadata records as inputs for Knowledge Graphs. Thus, they do not solve the initial issue of a costly manual curation of metadata records. Instead, a Knowledge Graph-based solution can be applied earlier in the workflow to support data stewardship by a uniform description of both social media collections and provenance information about the collection process.

We reuse existing open-source tools – and metadata they produce – to generate a Knowledge Graph, addressing interoperability issues and enabling data stewardship. Therefore we support users in the GLAM domain with basic IT understanding but limited technical skills [24]. Because we provide a workflow based on open source software and data models, independent of particular archiving use cases, we consider our solution sustainable. We analyzed existing social media harvesting tools to identify promising reuse candidates. Then we complemented selected tools with open source components to design a sustainable workflow driven by a Knowledge Graph: heterogeneous data are mapped to RDF, from which domain-specific metadata records are generated via queries. We validate our workflow by applying it on a social media archiving use case at Royal Library of Belgium (KBR). in which we created a Knowledge Graph based on harvested Twitter content, and generate MARC and EAD records.

Our contributions are (i) a comparative analysis of existing social media archiving tools, and (ii) a sustainable social media archiving workflow based on declarative RML mapping rules to generate Europeana Data Model and PREMIS-based [8] RDF from heterogeneous data sources, and metadata record generation based on reusable templates and Knowledge Graph queries. These open source resources as well as a full version of the comparison are available at https://github.com/RMLio/social-media-archiving.

In Section 2 we present related work. In Section 3 we provide a comparative analysis of social media harvesting tools. In Section 4 we present our Knowledge Graph-based

[2] Galleries, Libraries, Archives, and Museums.
[3] https://www.loc.gov/marc/bibliographic/
[4] https://www.loc.gov/ead/

solution which we validate in an archiving use case in Section 5. Finally, in Section 6 we discuss and conclude.

2. Related Work

To the best of our knowledge, there are no openly available workflows for social media archiving which cover both harvesting and cataloguing in an automated fashion. We discuss (i) tools and frameworks related to web archiving and social media harvesting in Section 2.1, to reflect on existing efforts to archive social media, (ii) metadata standards of the GLAM domain related to archiving in Section 2.2, to elaborate on domain-specific practices, and (iii) how our solutions compares to existing Knowledge Graph-based solutions in Section 2.3.

2.1. Social Media Archiving

We discuss web archiving, tools to harvest social media, as well as methodologies and tools used in the GLAM domain to analyze social media.

Commonly-used workflows for web archiving involve (i) describing collections, i.e. which website domains should be harvested and how often, (ii) fetching content using web harvesters, e.g., Heritrix [22] to preserve websites in Web ARChive (WARC) files [20], a format to preserve both content and HTTP requests, and (iii) accessing archived collections using replay software, e.g., WaybackMachine [27] or pyweb[5] as in the internet archive[6]. Software like Web Curator Tool [23] or Annotation and Curation Tool (w3act)[7] can be used as management interface to describe collections and schedule harvests. Websites for preservation are usually selected based on their top-level domain for which archival institutions may have a legal obligation to preserve its content. However, such workflows keep harvested information and metadata locked up in several data formats. Social media poses different challenges compared to web archiving due to its dynamic content [28] and different data formats used by different providers. Thus, web archiving workflows cannot be adjusted to sustainable social media harvesting workflows out of the box.

Similar tooling exists for social media archiving, but is limited to collection creation and harvesting. The modular frameworks Social Feed Manager (SFM) [15,21] and STACKS [17] create collections and schedule harvests. SFM reuses existing social media harvesters and wraps collections in WARC files, preserving harvested metadata while providing a uniform file format across harvested social media data. However, the replay of WARC files harvested in this way is difficult, because the content of the WARC files varies in format, i.e. harvested from different social media providers using different harvesting methods.

Social media can be harvested either by fetching data from Application Programming Interfaces (APIs) or via simulating a web browser. API-based tools, e.g., Twarc[8]

[5] https://github.com/webrecorder/pywb
[6] https://archive.org/
[7] https://github.com/ukwa/w3act
[8] https://github.com/DocNow/twarc

for Twitter or Instaloader[9] for Instagram, provide command line interfaces abstracting concrete API requests. They usually provide rich metadata represented as structured data. Tools like Brozzler[10] or Webrecorder/Conifer[11] harvest less metadata but preserve the look and feel. They simulate a browser or provide live recording functionality to harvest the HTML-based web version of social media content using the WARC format [20]. The aforementioned frameworks and tools create, describe and harvest social media collections. Technical details of API access are wrapped into user interfaces or command line tools, suitable for GLAM institutions with limited technical skills [24].

Several GLAM-related frameworks concern social media analysis related to social media harvesting, but not necessarily to social media archiving. In the case of ArchivesUnleashed[24], a project aiming to improve scholarly access to web archives, the collection development and harvests are explicitly excluded. Similarly, the GLAM workbench[12] aims for scholarly access by providing Jupyter notebooks[13], a combination of narrative text and live code. Candela et al. [4] investigated a methodology to create reproducible notebooks for the GLAM domain. Such frameworks are more concerned with analysis of already collected/described data and thus are complementary to our solution, i.e. they can be applied on archived data described with our Knowledge Graph.

2.2. *Metadata Standards and Cataloguing*

We discuss existing metadata standards and tools to create records adhering to those standards. The Online Computer Library Center (OCLC)[14], a global library cooperative, released recommendations for web archiving metadata fields [9]. They distilled 14 elements from the general vocabularies Dublin Core[15] and Schema.org[16], the XML-based standards Encoded Archival Description (EAD)[4], MARC21[3], and the Metadata Object Description Schema (MODS)[17].

However, the structure in which such elements are used is equally important, several subtly different standards exist. The General International Standard Archival Description (ISAD(G))[18] provides general guidance for the preparation of archival descriptions. EAD is a document-based hierarchical standard used to describe archival records. Although EAD is criticized to be document-centered rather than data-centered[13], hierarchical EAD records can be used to describe social media collections[19]. Compared to archival standards, MARC21 and MODS are bibliographic standards more focused on the library domain. The Metadata Encoding & Transmission Standard (METS)[20] encodes

[9] https://github.com/instaloader/instaloader
[10] https://github.com/internetarchive/brozzler
[11] https://github.com/Rhizome-Conifer/conifer
[12] https://glam-workbench.github.io/web-archives/
[13] https://jupyter.org/
[14] https://www.oclc.org/en/home.html
[15] https://dublincore.org/
[16] https://schema.org/
[17] http://www.loc.gov/standards/mods/
[18] https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition
[19] Collection of social media posts from Facebook and Twitter: https://tiaki.natlib.govt.nz/#details=ecatalogue.1016365 https://tiaki.natlib.govt.nz/#details=ecatalogue.1016484
[20] https://www.loc.gov/standards/mets/

descriptive, administrative, and structural metadata regarding objects within a digital library, popular to describe elements on an item level[7,10]. Incorporating all standards in a single model is difficult, as they take different perspectives [14]. Thus, we designed a Knowledge Graph in RDF, generated from heterogeneous born-digital data sources and described using domain-specific vocabularies. This allows generating records of different metadata standards.

Existing tools to generate archival metadata records are usually manual or semi-automatic cataloguing tools, closed source or commercial. According to embedded technical metadata, available EAD records for social media collections[19] are generated from the tool KE EMu[25]. Similarly, the ArchivesHub[21], a portal to integrate collections of several UK archives, uses the commercial software CIIM[22]. Such cataloguing tools are commercial software relying on existing archival records, either created manually or integrated from existing collections, and do not solve the problem of a costly manual creation. In our case, collection information is integrated via open source software from heterogeneous data sources and metadata records are generated automatically. Thus, web archivists are supported by initially generated metadata records to refine if necessary.

2.3. Knowledge Graph-based solutions

The GLAM domain already recognized Knowledge Graphs as promising future direction [2]. Dedicated ontologies and RDF representations for data models were developed, such as the official RDF ontology for MODS[23] and XSL Stylesheets to transform EAD documents to some RDF representation[24]. However, those RDF representations and ontologies do not describe data and their provenance, but metadata records summarizing data from a specific perspective.

The Europeana Data Model (EDM) [8], developed with technical experts from the GLAM domain, was designed to accommodate different standards. It represents a cultural heritage object together with different representations of it and contextual metadata. ArDO [30] is an ontology for hierarchical multimedia archival records based on specific application requirements and thus not extending EDM, but reusing it as guidance. Hierarchical archival data are also possible metadata records in our case. We use EDM and enrich our data with other more domain-specific vocabularies, e.g., TweetsKB [11] for social media content, and Dublin Core Collection Description[25] to describe social media collections. The PREMIS Data Dictionary for Preservation Metadata is a standard for which an ontology was developed [5], in version 2.2, meanwhile succeeded by a new ontology version to reflect PREMIS changes of version 3[26]. PREMIS was built on the Open Archival Information System (OAIS) reference model, an ISO standard [19] which among others describes different information packages. We reuse the PREMIS ontology to describe harvested data and its provenance. Similarly to EDM, PREMIS distinguishes between an actual object and its different representations, easing the integration with EDM and the rest of our model.

[21] https://archiveshub.jisc.ac.uk/

[22] https://www.k-int.com/products/ciim/

[23] https://www.loc.gov/standards/mods/modsrdf/

[24] http://data.archiveshub.ac.uk/ead2rdf/

[25] https://www.dublincore.org/specifications/dublin-core/collection-description/collection-application-profile/

[26] https://www.loc.gov/standards/premis/ontology/owl-version3.html

Tool	Approach	Output format	Social Media providers			Setup	Config	PROV
			T	F	I			
4CAT	Framework	JSON	+	-	+	advanced	UI	+
APIBlender	Framework	JSON	+	+	-	n/a	file	n/a
Brozzler	Browser	WARC/ HTML	+	+	+	advanced	file	+
Instaloader	API	JSON	-	-	+	beginner	file	+
DMI-TCAT	API	SQL	+	-	-	advanced	file	+
STACKS	Framework	JSON	+	-	-	advanced	file	+
SFM	Framework	WARC/ JSON	+	-	-	advanced	UI	++
Twarc	API	JSON	+	-	-	beginner	file	+
WebRecorder/ Conifer	Browser	WARC/ HTML	+	+	+	advanced	UI	+

Table 1. A comparison of features of different social media harvesting tools, T=Twitter, F=Facebook, I=Instagram. Full version available online https://github.com/RMLio/social-media-archiving

Regarding archival records, Knowledge Graph solutions are mostly applied on top of existing archival descriptions. Dobreski et al. [7] generate Linked Data for non-textual item-level data, e.g., images, sound, and videos, from XML-based archival records. Hennicke et al. [18] described how existing Bibliopolis and EAD records can be converted to EDM. Although only few Linked Data principles are followed, Gartner [13] devised a solution to represent archival description in a more constrained version of EAD as XML Schema from which regular EAD records can be generated. In contrast to these solutions, we do not generate a Knowledge Graph from existing metadata records and taking their perspective, but integrate raw data into a Knowledge Graph and generate different domain and perspective-specific metadata records in a following step. This way, we avoid the costly manual creation of archival records in the first place, while still providing means to curate data and metadata records.

3. Comparative Analysis of Social Media Harvesting Tools

Several social media archiving tools exist, varying in supported social media providers, usability and functionality. We compare available open source tools based on features relevant to social media archiving (Table 1).

We adapt a framework of the Data Together Initiative[27] originally used to compare generic web harvester tools. We reuse existing columns and add specific columns related to social media archiving in the GLAM domain. We compare the tested tools based on their approach, output format, setup, supported social media providers, configuration, and provenance. All tools but *APIBlender* are still maintained, i.e. commits or pull requests which indicate maintenance.

[27] https://github.com/datatogether/research/tree/master/web_archiving

Approach and output format The approach followed by the tool to harvest social media data and influences the output format: querying data from a single API, simulating a browser, or providing a whole framework. Despite their different approaches, all tools provide interfaces to abstract from the technical aspects of harvesting, and therefore have the potential to suit users in the GLAM-domain.

Different use cases demand different approaches. API-based tools provide machine-readable JSON data and can be used to harvest large amounts of data facilitating further analyses. Even though most JSON harvesting tools store data as files, STACKS stores JSON in a MongoDB and DMI-TCAT in a relational MySQL database. This may increase performance when interacting with the data, but in the case of MySQL also involves yet another data format negatively influencing interoperability. On the other hand, tools simulating a browser store HTML content in WARC containers and thus preserve the look and feel and performed HTTP requests, but usually are slower and may pose more technical challenges compared to API-harvesters as social media content is dynamic [28].

Frameworks provide harvesting functionality for several social media providers and graphical user interfaces, and a promising code base for GLAM institutions. They are usually extensible with own modules or use existing harvesters, e.g. SFM uses Twarc for Twitter harvests. The output format for such frameworks usually depends on the harvesters used, but interestingly, SFM harvests data in JSON format, but preserves it in WARC files [21]. Thus, it provides a uniform interface of harvested social media data across providers while preserving technical metadata which positively influences downstream tasks requiring provenance.

Supported social media providers From which social media providers the tool can harvest data. For this analysis we consider Twitter, Instagram and Facebook as they are part of the long-term goals for our BeSocial use case. Most tools support Twitter, some Instagram and only a few Facebook. Tools harvesting Facebook are simulated browsers, technological challenges for Facebook might be a reason [29] why no tool uses other harvesting means for Facebook. API-based tools are focused on a single provider, frameworks usually support several providers, and tools simulating a browser are technically not limited to any provider as they aim to harvest web content in general. Therefore, either frameworks or simulated browser tools are promising candidates if several social media providers should be supported by the use case.

Setup of the tool We distinguish two levels of difficulty for setting up tools for harvesting: *beginner*, where only a script needs to be installed using a package manager; *advanced*, where several components need to be installed. Most tools can be set up with minimum programming experience, e.g., only by installing one command line tool. The majority of tools requires more steps as they consist of several components. However, such tools usually provide means to compensate, e.g. by providing docker images which, can be started and stopped as containers with minor configuration and a single command, or by providing the harvester as a service. Yet, debugging of such a docker setup, if needed, requires a deeper technical understanding, possibly challenging for users in the GLAM domain.

Tool's configuration How the tools can be used to create social media collections: the more technical abstractions, the better considering less-technical users. All tools are configured via config files or web interfaces, lowering the reuse barrier.

Provenance information Technical metadata captured via the harvesting process and/or descriptive metadata from the harvested content, considering archiving: usually the more the better. In terms of harvested content, tools harvesting data from APIs usually provide rich descriptive metadata facilitating analyses and data stewardship tasks, whereas tools harvesting HTML content in WARC files only provide technical metadata within the WARC HTTP headers. From a collection-level point of view, descriptive metadata in form of collection description needs to be added manually via the configuration of the tools. Regarding provenance information, SFM provides the best trade-off as it preserves technical metadata from harvests within WARC files, descriptive metadata of harvested content as part of the API responses, and descriptive metadata of collections – entered by users via a UI – within a relational database.

Discussion Since frameworks may reuse existing harvesters, they are promising reuse candidates for use cases where several social media providers are considered for archiving. Compared to other frameworks, SFM has the advantage of storing harvested data within WARC files which provides additional provenance information. Additionally, collections in SFM are configured via an user interface which addresses users of the GLAM domain and thus our use case.

4. Sustainable Workflow

Our workflow reused open-source components to (i) describe social media collections, (ii) harvest social media content, and generate (iii) Knowledge Graphs and (iv) domain-specific metadata records. We present our modular architecture (Fig. 1) based on open source frameworks in Section 4.1 and discuss design decisions regarding regarding RDF representations in Section 4.2.

4.1. Architecture and Components

Our modular solution integrates into an existing framework and provides three declarative ways to control the social media archiving workflow. We describe the components of our architecture with the following contributions: (i) integration of automatic Knowledge Graph generation into the existing social media harvesting framework SFM, (ii) reusable declarative Knowledge Graph generation rules to describe social media archives, and (iii) reusable declarative queries and templates to generate domain-specific metadata records.

Social media harvesting We reuse the Social Feed Manager (SFM) where a central RabbitMQ message queue is used for communication among components. Archivists create social media collections via a UI where they specify the seeds to harvest, a harvesting schedule, and provenance information regarding the collection (Fig. 1, 1), i.e. title and description. At specified intervals a harvesting message is sent to the message queue which triggers existing social media harvesters, e.g., Twarc for Twitter, to fetch data.

SFM supports several API-based harvesters and uses a WARC proxy to preserve technical provenance information by recording performed HTTP requests and store them together with the received HTTP response in WARC files (Fig. 1, 2). Thus, SFM offers a uniform file format with technical provenance information for differently described so-

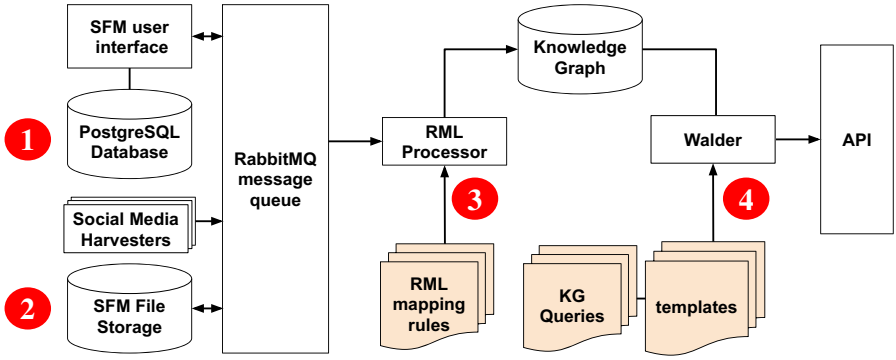

Figure 1. Our sustainable social media archiving workflow's architecture is based on open source components and is controlled with only three lightweight and declarative components (orange): RML mapping rules to create Knowledge Graphs, templates to specify metadata records, and queries to populate the templates.

cial media content from different social media providers. We utilize this uniform format to generate interoperable provenance across social media content. Harvesters indicate the status to the message queue, e.g., a successful harvest with listed information such as the location of newly created WARC files, which we use as input for our Knowledge Graph generation.

Knowledge Graph generation SFM provides a rich source of heterogeneous (meta) data which we lift to a Knowledge Graph to get a uniform and interoperable description of captured and preserved social media. We integrate descriptive collection metadata from SFM and the content harvested, as well as technical metadata produced by SFM and enclosed in preserved WARC files.

We use the RML.io framework[28] (Fig. 1, 3) to generate the BESOCIAL Knowledge Graph. RML.io generalizes the W3C recommended R2RML specification [6] to integrate heterogeneous data based on declarative mapping rules which is needed for our use case. We use the RMLMapper[29] to generate the Knowledge Graph based on declarative mapping rules following the RML specification.

Metadata records generation Although a Knowledge Graph-based data model enables semantic interoperability of data, concrete preservation systems or other stakeholders in the GLAM domain demand metadata records summarizing certain data in a domain-specific syntax, e.g. MARC21 for libraries, EAD for archives. We provide a component to automatically generate such metadata records from our Knowledge Graph avoiding a costly manual curation. We use Walder[30] which allows setting up a website or API over decentralized knowledge graphs. Using existing template libraries from web development, e.g., Handlebars[31], templates for metadata records are created. The query language GraphQL-LD [26] is used to query the Knowledge Graph and populate declarative templates with content, generating metadata records published via an API using Walder (Fig. 1, 4), while avoiding needing in-depth programming experience.

[28] https://rml.io
[29] https://github.com/RMLio/rmlmapper-java
[30] https://github.com/KNowledgeOnWebScale/walder
[31] https://handlebarsjs.com/

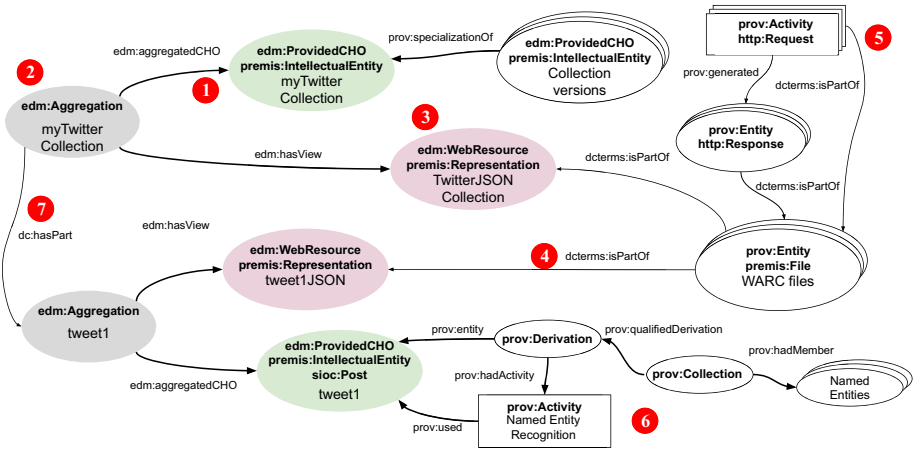

Figure 2. The Europeana Data Model (EDM) is used to represent social media collections and posts as cultural heritage objects (green) and their different representations (violet), aligned with PREMIS and PROV to represent provenance.

4.2. Data-driven Workflow

We describe how the Europeana Data Model (EDM), the de-facto standard for cultural heritage data, and other common W3C recommended vocabularies can be used to represent social media collections in an interoperable way.

We followed a Competency Question (CQ)-based approach, commonly used to express requirements in ontology engineering [16]. We defined more than 20 CQs for our archival use case based on user-stories to determine which data needs to be integrated. A full list is openly available at our online resource.

We reuse the EDM to describe harvested social media content because it enables us to represent not only the object itself, e.g. a Tweet via its ID, but also differently harvested representations, e.g. captured JSON or HTML representations of a Tweet stored in WARC files. A whole collection, created by users via SFM and stored in a relational database, and social media posts (items of the collection) are represented as cultural heritage objects using the class `edm:ProvidedCHO` and `premis:IntellectualEntity` (Fig. 2, 1). Such a collection or item may have different representations linked by an instance of `edm:Aggregation` (Fig. 2, 2), in our case the harvesters used by SFM fetch information in JSON from APIs, and thus we use `edm:WebResource` and `premis:Representation` to represent a JSON representation (Fig. 2, 3); someone may harvest social media posts (additionally) in their HTML representation which would then be another `edm:WebResource`, linked to the associated aggregation (Fig. 2, 2). To increase interoperability we represent social media posts also using `sioc:Post` from TweetsKB [11].

Harvested social media data is enclosed in WARC files by SFM (Fig. 2, 4) preserving harvest metadata of HTTP requests. We represent such harvest metadata using PROV activities, listing when and how WARC files were created (Fig. 2, 5), WARC files are represented using `premis:File`. On item level, we perform Named Entity Recognition

(NER) during mapping via the DBpedia spotlight API[32] to enrich our Knowledge Graph (Fig. 2, 6). This information is useful later when generating archival records. PROV is used to preserve information of the NER process. Hierarchical information, such as which item belongs to which collection, is explicitly represented using Dublin Core and following EDM guidelines[33] (Fig. 2, 7).

5. Social Media Archiving at KBR

BESOCIAL is a cross-institutional research project, aiming to develop a sustainable strategy for archiving and preserving social media in Belgium. The solution supports this goal by offering a sustainable social media archiving workflow. We outline the use case and describe how we applied our workflow within a pilot.

BESOCIAL use case KBR, as the federal scientific library of Belgium, is legally mandated to collect and preserve all Belgian publications. To tackle challenges of the digital-era, KBR invests in the digital preservation of online content. In the past KBR worked on a federal strategy for the preservation of the Belgian Web [28]. Due to the uniqueness and ephemeral nature of social media, BESOCIAL brings together interdisciplinary partners to consider conservation, preservation and accessibility of developing a social media archive.[34] Twitter was selected as promising social media platform, but Instagram and Facebook are considered in the long-term. Recent outcomes of BESOCIAL are the analysis of an online survey in which 15 international archiving institutions participated, and which showed that many institutions are engaged in social media archiving, but also that the stage and efforts vary in size and scope [29].

Content selection Web archivists define so-called seed lists with content that should be archived. For BESOCIAL, a seed list with 86 relevant Belgian entities of 14 categories, such as governmental institutions and online news, was curated by KBR for a test pilot. From these 86 entities, 79 had accounts on Twitter. We used the user interface of SFM to create a collection for these accounts.

Content collection Collections created with the user interface of SFM were scheduled to harvest social media data daily. This, so far, resulted in 50 compressed WARC files of 88 MB enclosing around 200,000 Tweets in JSON format. The first harvest resulted in roughly 150,000 tweets as the used Twarc harvester of SFM fetches the most recent 3,200 tweets per account. Subsequent daily harvests resulted in less content of up to 2,000 tweets. These are heterogeneous data which we need to lift to a Knowledge Graph to facilitate data stewardship tasks.

Knowledge Graph generation We used the data model and its requirements expressed as Competency Questions (CQs) described in Section 4.2 to systematically guide the integration process, i.e. one RML mapping contributes data to answer at least one CQ. Applying these mappings resulted in one RDF file per WARC file and one RDF file for collection-level metadata extracted from the SFM PostgreSQL database. We generated

[32] https://www.dbpedia-spotlight.org/api

[33] https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Mapping_Guidelines_v2.4_102017.pdf

[34] https://kbr.be/en/projects/besocial

RDF triples consisting among others of 213,000 EDM cultural heritage object resources representing collections and social media posts, and 222,000 W3C PROV activities reflecting provenance.

Metadata records generation Different domain-specific data formats exist. Already available social media collections are described using EAD records[19], thus we consider this a baseline, and KBR as a library works with bibliographic MARC-based records to describe collections. Additionally, human users may want to browse collections. Thus, we created two XML-based and one HTML-based template and related GraphQL queries for Walder to populate these templates from our Knowledge Graph to accommodate these use cases; available at our online resource[35]. We can query heterogeneous data, to among others, get aggregated information about named entities, enabling users to assess the content i.e. which locations or events are mentioned within a whole social media collection. Hierarchical information is present in our Knowledge Graph as we reused terms like `dc:hasPart` (Fig. 2, 7).

Discussion We discuss the added value of the Knowledge Graph in our use case and findings related to the Knowledge Graph's use with respect to collection-level and item-level (social media post) data.

Instead of many-to-many mappings from heterogeneous data sources to heterogeneous metadata records, our solution results in a semantically described RDF Knowledge Graph which facilitates data stewardship as it describes all preserved data including provenance information. The generation of metadata records and HTML views are thus not limited to harvested data, but also profit from contextual information of the Knowledge Graph, because item-level data (social media posts) are put in relationship to collections and provenance information. This information can be queried using SPARQL or GraphQL, therefore we are able to identify e.g. social media posts belonging to different collections or collections/posts mentioning similar named entities. Similarly, more fine-grained queries are possible with more integrated linked data in the future, i.e. archivists may rather spend manual curation efforts in enriching the Knowledge Graph instead of domain-specific metadata records.

Use cases related to the collection-level may not need the full graph. Whereas harvested data preserved and compressed in WARC files are relatively small, the Knowledge Graph is considerably larger. This may present a performance bottleneck for smaller setups without adequate RDF database or hardware. However, HTML views providing an overview of collections, or MARC records describing bibliographic information of collections do not need all item-level details such as detailed post provenance. We used decentralized Knowledge Graphs partitioned between collection and item level data to improve performance of collection-level tasks.

If certain use cases demand some item-level information we declaratively create aggregations. Based on the data model and extracted information, we used SPARQL-CONSTRUCT queries to enrich collection-level information with aggregated information from item-level, such as most often used named entities and their type; vocabularies such as the W3C recommended WebAnnotations[36] or DataCube[37] may be used to semantically describe aggregates, further research is required.

[35] https://github.com/RMLio/social-media-archiving
[36] https://www.w3.org/TR/annotation-model/
[37] https://www.w3.org/TR/vocab-data-cube/

Libraries usually provide full access to collections only via reading rooms or after login, and from a legal perspective it is also problematic to provide public access to harvested social media data. However, collection-level related parts of the Knowledge Graph including aggregations present a smaller sub-graph which may be made publicly available, directly as API or via HTML views. Therefore, end users may assess more detailed information about collections using contextual-rich collection information before requesting access to the full collection on-premise or online which could positively influence the user experience. However, more research towards the needs of different types of users is needed.

6. Conclusion

Social media is already a paramount part of our society and, thus, its content needs to be preserved. However, archiving is an expensive long-term commitment and currently only fragmented workflows for social media archiving exists. We developed an open source Knowledge Graph-based solution using the Europeana Data Model and PREMIS to describe WARC-preserved social media as cultural heritage objects with different representations. Now we can support automatic generation of GLAM-related metadata records, e.g., MARC and EAD, or provide collection overviews via HTML for users to assess the collections' content.

Human-in-the-loop provenance Social media harvesting tools play a crucial role regarding provenance information, as they cover initial phases of selection and collection where human users define what to harvest and when. Currently SFM provides a detailed change history of collections, but descriptive information is limited to titles and descriptions. Similar to how some web archiving tools require the upload of legal deposit documents before harvests are initiated [23], SFM could be extended with UI fields to collect specific information from users in a uniform fashion. Our Knowledge Graph-based solution allows a data-centric perspective driven by downstream tasks which can inform improvements of SFM's UI and database, to include more, and more-specific metadata fields which would positively influence the quality of generated metadata records.

Data stewardship of digital collections Social media archives are not static and pose new challenges for which data stewardship is needed: some content may have to be removed from public access due to intellectual property or privacy-related take-down requests, and on top of that several terms of services from different social media providers need to be taken into account. Such stewardship tasks are supported by our solution. For example, our Knowledge Graph already encodes provenance information of harvesting, and as it is based on PREMIS and W3C PROV, existing data can be annotated or additional provenance information regarding take-down requests can be included in the same fashion. Therefore, consuming applications can perform policy-compliant operations with the harvested data.

Future Work Future work will investigate the quality of generated metadata records and extend the metadata record queries if necessary. The modular tool SFM can be extended with new functionality or other social media harvesters. Based on our Knowledge Graph, operational and legal challenges of social media archiving can be reconsidered and addressed.

Acknowledgements The research activities were supported by the Belgian Federal Science Policy Office (BELSPO) BRAIN 2.0 Research Project BESOCIAL, Ghent University, imec, and Flanders Innovation & Entrepreneurship (VLAIO).

References

[1] Acker, A., Kreisberg, A.: Social Media Data Archives in an API-driven World. Archival Science **20**(2), 105–123 (2020)
[2] Bahnemann, G., Carroll, M., Clough, P., Einaudi, M., Ewing, C., Mixter, J., Roy, J., Tomren, H., Washburn, B., Williams, E.: Transforming metadata into linked data to improve digital collection discoverability (2021)
[3] Borgman, C.L.: Scholarship in the digital age: Information, infrastructure, and the Internet. MIT press (2010)
[4] Candela, G., Sáez, M.D., Escobar Esteban, M., Marco-Such, M.: Reusing digital collections from GLAM institutions. Journal of Information Science (2020)
[5] Coppens, S., Verborgh, R., Peyrard, S., Ford, K., Creighton, T., Guenther, R., Mannens, E., Van de Walle, R.: PREMIS OWL. International Journal on Digital Libraries **15**(2), 87–101 (2015)
[6] Das, S., Sundara, S., Cyganiak, R.: R2RML: RDB to RDF Mapping Language. Working group recommendation, World Wide Web Consortium (W3C) (Sep 2012), http://www.w3.org/TR/r2rml/
[7] Dobreski, B., Park, J., Leathers, A., Qin, J.: Remodeling archival metadata descriptions for linked archives. In: International Conference on Dublin Core and Metadata Applications. pp. 1–11 (2020)
[8] Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C., Van de Sompel, H.: The Europeana Data Model (EDM). In: World Library and Information Congress: 76th IFLA general conference and assembly. vol. 10, p. 15 (2010)
[9] Dooley, J.M., Bowers, K.: Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group. OCLC Research (2018)
[10] Elings, M.W., Waibel, G.: Metadata for all: Descriptive standards and metadata sharing across libraries, archives and museums. First Monday (2007)
[11] Fafalios, P., Iosifidis, V., Ntoutsi, E., Dietze, S.: TweetsKB: A Public and Large-Scale RDF Corpus of Annotated Tweets. In: European Semantic Web Conference. pp. 177–190 (2018)
[12] Fondren, E., McCune, M.M.: Archiving and Preserving Social Media at the Library of Congress: Institutional and Cultural Challenges to Build a Twitter Archive. Preservation, Digital Technology & Culture **47**(2), 33–44 (2018)
[13] Gartner, R.: An XML schema for enhancing the semantic interoperability of archival description. Archival Science **15**(3), 295–313 (2015)
[14] Gartner, R., Mouren, R.: Archives, museums and libraries: breaking the metadata silos. In: Paper presented at IFLA WLIC 2019. Athens, Greece (2019)
[15] George Washington University Libraries: Social feed manager. version 2.3.0 (May 2020). https://doi.org/10.5281/zenodo.3784836
[16] Grüninger, M., Fox, M.S.: The Role of Competency Questions in Enterprise Engineering. In: BenchmarkingTheory and practice, pp. 22–31. Springer (1995)
[17] Hemsley, J., Jackson, S., Tanupabrungsun, S., Ceskavich, B.: STACKS - Social Media Tracker, Analyzer, & Collector Toolkit at Syracuse (Apr 2019). https://doi.org/10.5281/zenodo.2638848
[18] Hennicke, S., Olensky, M., de Boer, V., Isaac, A., Wielemaker, J.: A data model for cross-domain data representation. In: Proceedings of the 12th International Symposium on Information Science. pp. 136–147 (2011)
[19] ISO Central Secretary: ISO 14721:2012 Space data and information transfer systems. Standard ISO 14721:2012, International Organization for Standardization, Geneva, CH (2012)
[20] ISO Central Secretary: ISO 28500:2017 Information and documentation WARC file format. Standard ISO 28500:2017, International Organization for Standardization, Geneva, CH (2017)
[21] Littman, J., Chudnov, D., Kerchner, D., Peterson, C., Tan, Y., Trent, R., Vij, R., Wrubel, L.: API-based social media collecting as a form of web archiving. International Journal on Digital Libraries **19**(1), 21–38 (2018)
[22] Mohr, G., Stack, M., Rnitovic, I., Avery, D., Kimpton, M.: Introduction to Heritrix. In: 4th International Web Archiving Workshop. pp. 109–115 (2004)

[23] Paynter, G., Joe, S., Lala, V., Lee, G.: A year of Selective Web Archiving with the Web Curator at the National Library of New Zealand. D-Lib Magazine **14**(5/6), 1082–9873 (2008)
[24] Ruest, N., Lin, J., Milligan, I., Fritz, S.: The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020. pp. 157–166 (2020)
[25] Sendino, M.C.: KE EMu and the future for natural history collections. Collections **5**(2), 149–158 (2009)
[26] Taelman, R., Vander Sande, M., Verborgh, R.: GraphQL-LD: Linked Data Querying with GraphQL (2018)
[27] Tofel, B.: 'Wayback' for Accessing Web Archives. In: Proceedings of the 7th International Web Archiving Workshop. pp. 27–37 (2007)
[28] Vlassenroot, E., Chambers, S., Di Pretoro, E., Geeraert, F., Haesendonck, G., Michel, A., Mechant, P.: Web archives as a data resource for digital scholars. International Journal of Digital Humanities **1**(1), 85–111 (2019)
[29] Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J.: Web-archiving and social media: an exploratory analysis [to be published]. International Journal of Digital Humanities (2021)
[30] Vsesviatska, O., Tietz, T., Hoppe, F., Sprau, M., Meyer, N., Dessi, D., Sack, H.: ArDO: An Ontology to Describe the Dynamics of Multimedia Archival Records [to be published]. In: ACM, Symposium On Applied Computing (2021)