DATASET PRODUCTION AS A NEW PROCESS IN FUTURE AI-EMPOWERED MEDIA

Alberto Messina

RAI – Radiotelevisione Italiana, Italy

ABSTRACT

The current technological development trend in AI suggests that it will be a pervasive end-to-end component in all future media systems from production to distribution and considered as one of the “new normals” of a typical media production infrastructure of the near future. In this context, key assets of main European and world broadcasters are represented by the immense (and growing) number of archived objects, that together with their relevant metadata are seen by all major technology providers as a ground truth treasure. But how much is this belief true? Is it really so advantageous to consider archive metadata as an easy-to-use source of ground truth for machine learning tools? This paper will present the main challenges behind this approach and how these could be addressed by applying a rigorous and structured approach at what can be identified as a new process: dataset production. By defining and following key requirements for the dataset production process, the paper will illustrate some basic tools enabling decision taking about the effectiveness of the possible alternatives (e.g., metadata adaptation vs. metadata re-make) and will propose a theoretical background for the generation of future-proof datasets.

INTRODUCTION

In the current era the usage of Artificial Intelligence (AI) technologies in industrial processes is becoming commonplace in many sectors including finance, manufacturing, automotive and – of course – media and entertainment. The applications range is extremely wide and goes from business data analytics to automated quality control, web & social mining, multimedia classification, automated driving and many more (6)(7). The level of penetration of AI tools in production processes is evolving from a simple support to business decisions to full-fledged substitution of human decision makers. If on the one side this scenario poses unprecedented challenges in terms of ethics, labour policy, safety and liability, on the other side represents an unmissable opportunity to implement new areas of business otherwise unfeasible. The media sector is certainly one in which the potential of AI technologies may give its best results, and probably one in which risks related to the application of AI in the value chain can be better-mitigated w.r.t. other critical sectors (e.g., healthcare, finance, automotive) due to its inherent nature. There is also another key enabling factor in the media

domain, namely the availability of an alleged immense amount of data. However, how much of this data is actually usable, through what processes, and at what cost? This paper tries to elaborate on this problem, based on the observations and experiences of the past 20+ years in applied R&D in the field of AI in media processes (especially in archives).

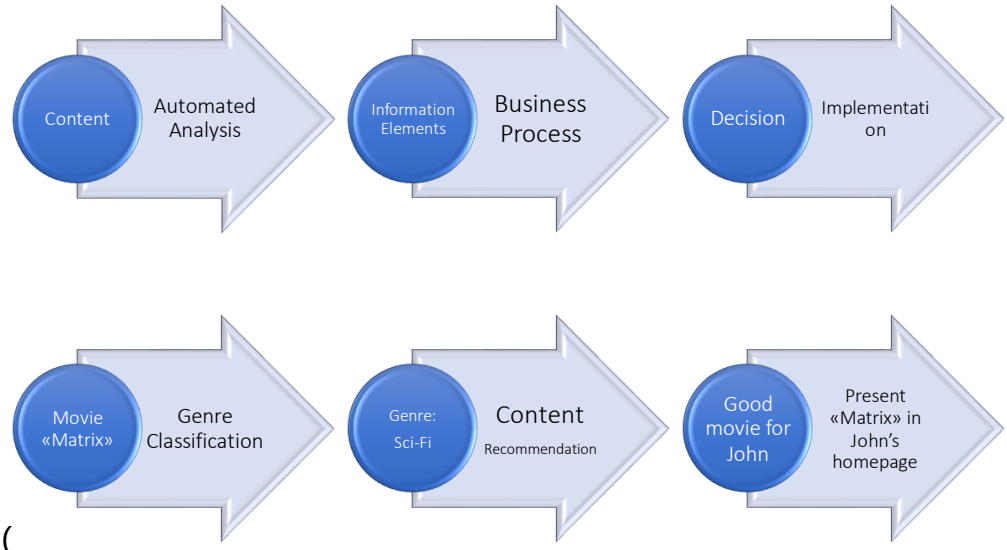
AI AND ARCHIVES: A LONG STORY

The application of AI in the realm of broadcast and media archives dates back more than two decades by now. First European R&D collaborative projects around this topic had been launched since the late nineties and counted automation of archive documentation among their strategic objectives[1][2]. Until the recent past, the classical approach has been to use available technologies and tools trained externally to annotate content, with few examples exploiting available documentation to train AI models. This last approach has been revitalised in recent years due to the explosion of advanced machine learning approaches based on deep neural networks, becoming affordable thanks to the increase of computation power density and stable integration of GPU technologies as part of the development stack (8). These modern approaches need considerable amount of data to work properly: in fact, different from legacy machine learning, they converge on the extraction of the most promising features through iterative learning rather than working on features engineered in advance. For popular applications in the media domain like image/video classification, technology providers see the availability of archive documentation as a source of unlimited (and typically low-cost) ground truth.

Ground Truth Generation is a Multi-faceted Problem

Many tasks which can benefit from AI support are characterised by a common pattern, namely that of extracting one or more information elements from the automated analysis of content. Information elements are part of a description model, which represents the way in which content is described according to the conceptual scheme characterising each specific domain. Information elements are normally used in some business processes to make

[1] https://cordis.europa.eu/project/id/FP4_24956/it
[2] https://cordis.europa.eu/project/id/IST-1999-20013



decisions (Figure 1).

In order to work as expected, a tool performing the mentioned analysis must be trained to identify and extract the target information elements through learning by certified examples of the association of such elements with certain content items (a.k.a. ground truth). Instances of such associations are given the name of datasets in the machine learning jargon. Since datasets for real-life tasks are scarce and expensive to produce, a common idea to solve this problem is to exploit available archive documentation as a source of ground truth and hence of datasets. However, if this approach may have proved successful when remaining in the same domain (e.g. train on archive data to extract archive-related information elements), the cross-domain generalisation remains an unsolved problem. As an example, consider the way in which the same movie is classified in different databases (Figure 2). The three exemplified platforms respond to three different business cases, characterised by different business processes and decisions, and different kinds of users, hence the need to use different description

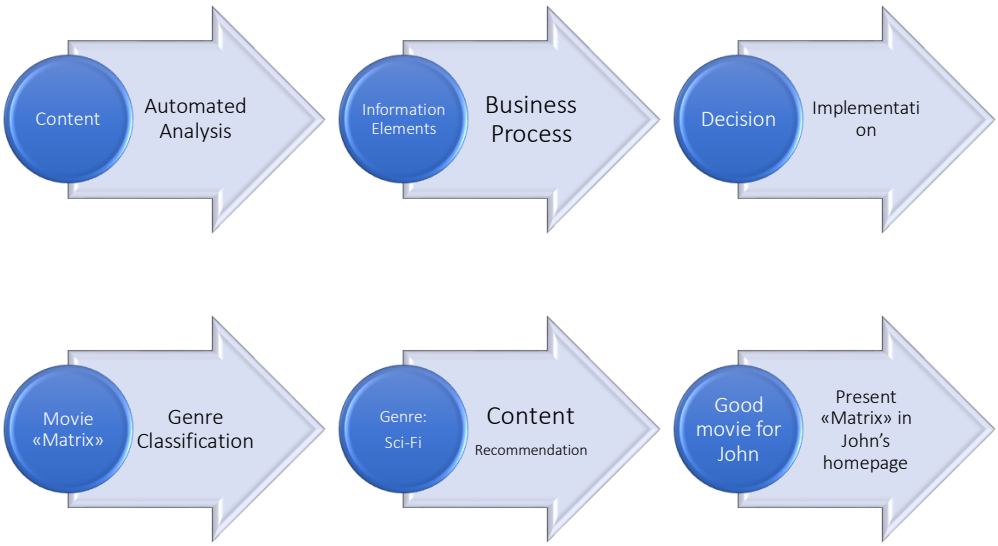
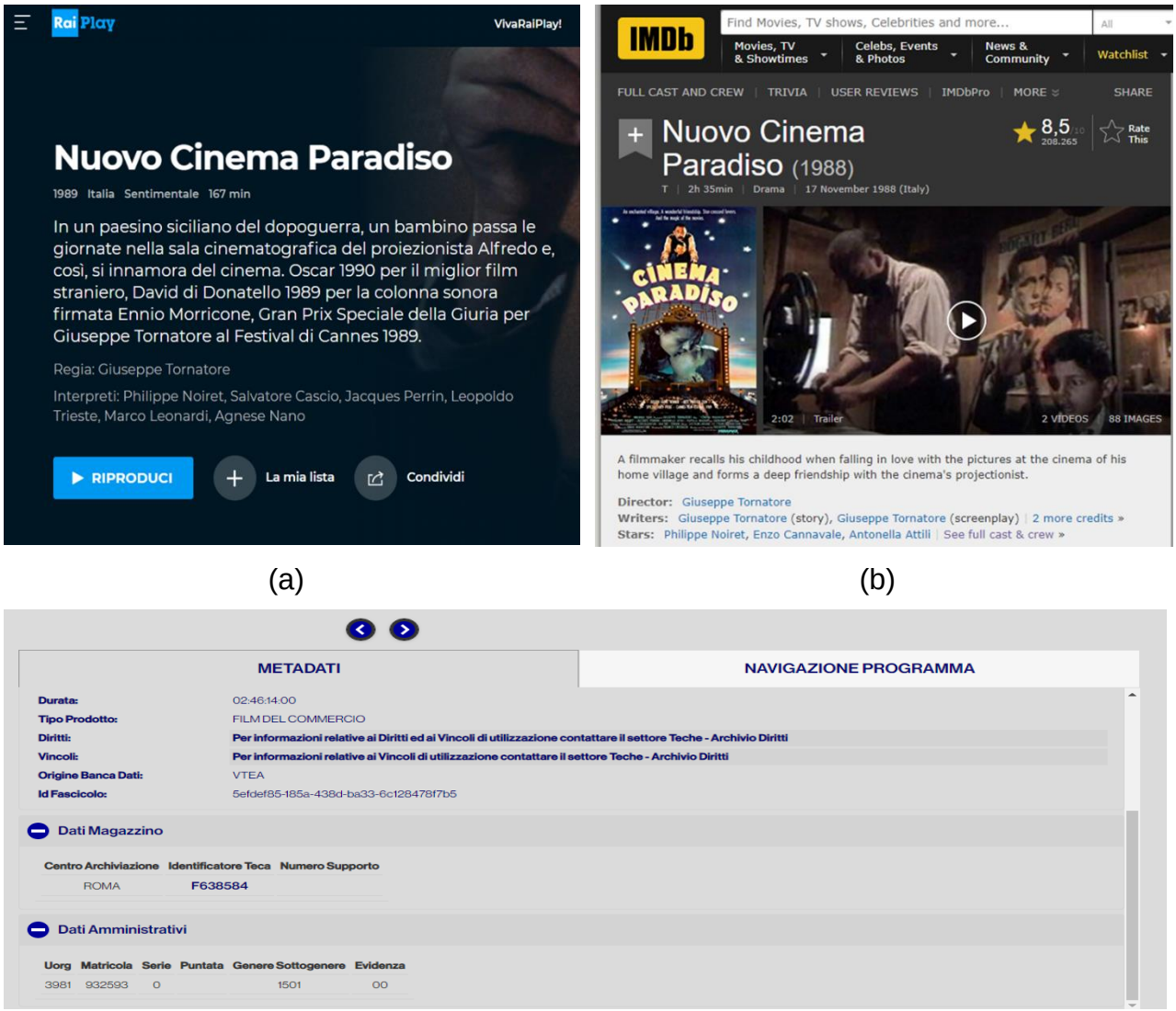


Figure 1. Information elements as inputs to business processes.

models. In such a scenario, an approach that trains AI tools on one domains' data has low likelihood of producing useful information in the others. In the specific case, a process trained on archive data (c) aimed at producing information elements useful for the OTT platform (a) would need that a thorough (and lossy) adaptation of description models is performed beforehand.



(a) (b)

(c)

Figure 2. Different annotations for the same movie: (a) from Rai Play; (b) from IMDB; (c) from RAI's archive catalogue.

We can therefore come to a first finding, i.e. that the relevance of information elements for a certain decisional task in a business process depends on the business domain, which means – ultimately – on the users taking part in the decision process. In the archive domain, the documentalists annotate content according to a well-defined description model, which

contains elements pertaining to decisions taken by users of the archives (e.g. whether or not the archived piece is useful for a new production, or whether appropriate rights are owned, or whether the media asset is of sufficient quality). In the OTT publishing domain editors annotate content according to a different description model, which contains elements helping final users take decisions about their interest for a programme. Although a full characterisation of the different description models is out of scope of this paper, for the sake of this discussion we can consider that description models can be represented along a three-dimensional space as depicted in Figure 3.

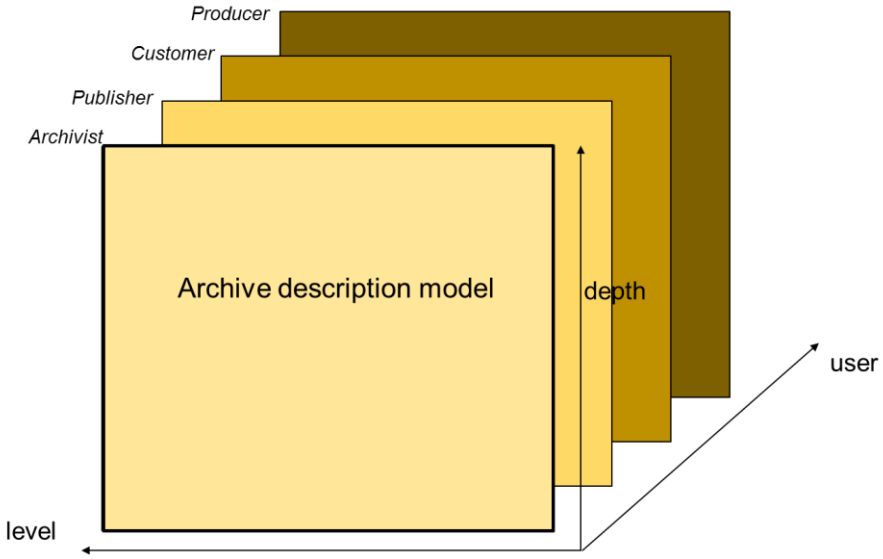


Figure 3. Description models representation.

Each user category corresponds to a bidimensional map and in each map, level and depth account respectively for the kind of information and the amount of detail for that kind. An example of a general depth/level map is depicted in Figure 4. Under such a simplified hypothesis, we could argue that each domain corresponds to a subset of the general depth/level map and subsets can have partial overlaps. Thus, the Archive description model can e.g. share part of the Stories sub-model with the Publisher description model and the Customer description model can share part of the Real World level with the Publisher. This establishes “information flows” across domains that, in principle, enable maximising cross-domain reuse of available data.

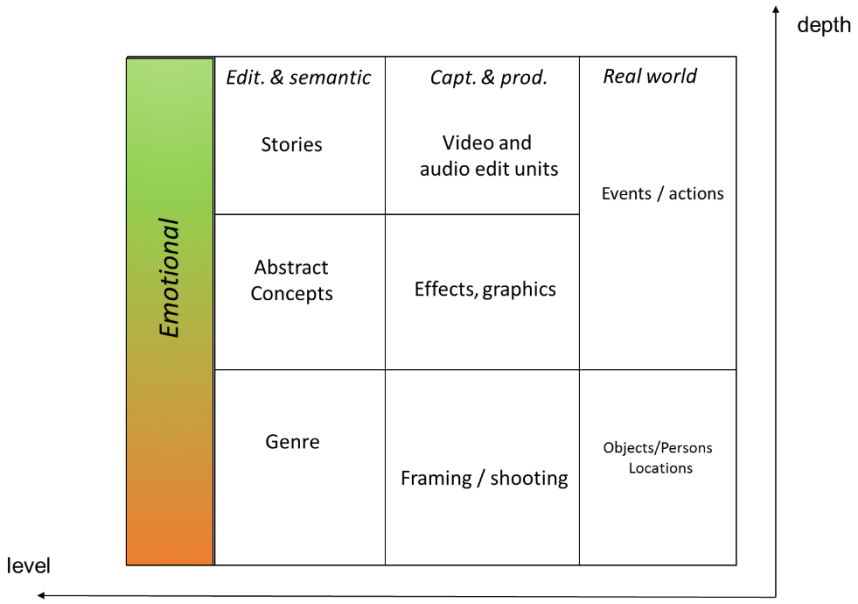


Figure 4. A possible general description model breakdown in depth/level.

However, do commonalities at the description model level ensure reuse of data for AI in the form of datasets? Let us also consider that business domain mismatch is not the only issue related to an archive-centred approach for ground truth generation. In fact, especially for broadcasters having deep archives other important issues are: 1) documentation data is stratified over many decades and compliant to different description models evolving over time; 2) long to mid-term variations of documentation budget can influence the depth and detail of annotation as well, resulting in heterogeneous data even for the same content genre; 3) different documentalists (users) may interpret and apply annotation criteria differently; 4) finally yet importantly, even in the case in which Information Elements are shared among business domains, they can be instantiated following different criteria[3].

DATASET PRODUCTION FUNDAMENTALS

The above discussion challenges basic criteria of machine learning that put homogeneity and statistical representativeness of data as key for proper functioning, so that any AI machinery targeted at extracting information elements in a real media environment must be designed according to the above fundamental findings. At the core of this design, there is a fundamental activity, that we can call dataset production.

Summarising, we can say that: 1) dataset production must comply to a business process – driven description model and is equivalent to the generation of information elements by the actors (users) of the business domain; 2) the way in which information elements are generated depend on the users, their context and modality of observation, and on other factors generally affecting their cognitive processes (e.g., attention, memory, culture) (See also (1)). In the remainder of this Section, we will formalise these concepts.

[3] Following the example depicted in Figure 2, IMDB's annotations for credits include names of actors appearing on the screen of Cinema Paradiso (e.g., Jean Gabin), while this information is not present in either of the other two cases.

Tensor-like Representation of Datasets

Let $BD = \{BD_1, BD_2, \dots BD_N\}$ be the set of *N* Business Domains and $U = \{U_1, U_2, \dots U_N\}$ the set of corresponding user classes, so that we denote with U_i the user class of BD_i. Let $DM = \{DM_1, DM_2, \dots DM_N\}$ be the set of description models for each Business Domain. We have that $DM_i = \{IE_{i1}, IE_{i2}, \dots IE_{iK_i}\}$, where IE_{ik} is the *k*-th Information Element of the description model DM_i. We can assume, as already anticipated, that description models may overlap, i.e. that $\exists p, q: DM_p \cap DM_q \neq \emptyset$. An Information Element can be e.g. uniquely identified through the specification of the URN of a metadata schema and of a set of criteria (annotation rules) $AR_i = \{R_{i1}, R_{i2}, \dots R_{iK_i}\}$ for the creation of data instances. Let us also denote with $Ext(IE_{ik})$ a generic set of instances of IE_{ik} generated according to these rules.

Let *M* be the set of media items of interest, which is shared among the Business Domains.

The dataset production process in each domain DM_i can then be denoted as a process function DP_i executed by a user $u_j \in U_i$ that takes *M* as input set and produces for each media item m_k the set of instances $Ext(DM_i)^{jk} = \{Ext(IE_{i1}), Ext(IE_{i2}), \dots, Ext(IE_{iK_i})\}^{jk}$ according to each Information Element's schema and generation rules (Figure 5).

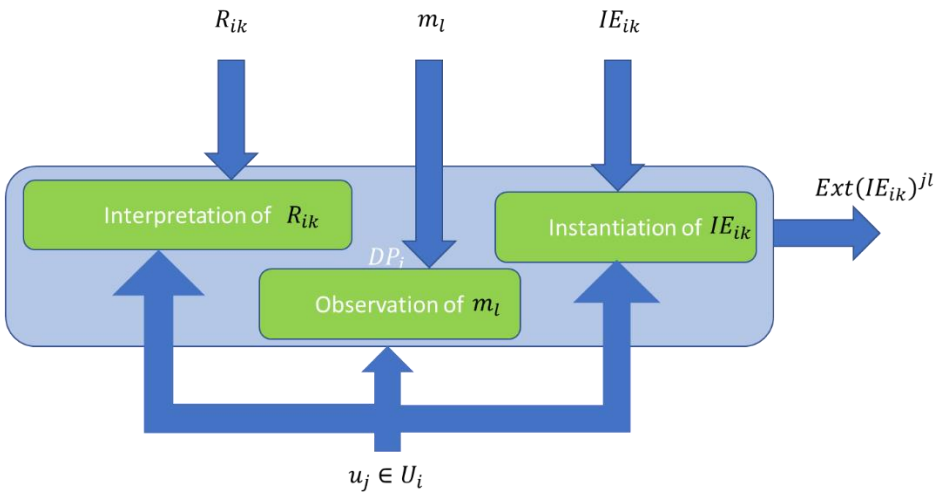


Figure 5. Formalised dataset production atomic process.

Analysing Figure 5 we can derive a formal definition of what can be called atomic ground truth statement, namely an Instance of Information Element IE_{ik} associated to a media item m_l and produced by user $u_j \in U_i$ according to his interpretation of production rule R_{ik}. The same schema also enlightens how dataset production is highly user-dependent (1).

In the most general terms, the result of the dataset production process in a Business Domain DM_i can be therefore represented as a 3D structure DS_i^{jkl}, where $j = 1 \dots |U_i|, k = 1 \dots |M|$, $l = 1 \dots K_i$ and $DS_i^{jkl} = Ext(IE_{il})^{jk}$.

As a simple example, consider a classical archive annotation process in which one documentalist ($|U| = 1$) annotates a collection *M* of items following a Documentation Model DM. For the sake of conciseness, let us assume that the documentation model includes recognition of Persons (IE_1) and Locations (IE_2) only ($|DM| = 2$). The resulting dataset is then of dimensionality $1x|M|x2$ (Figure 6 – (a)). If the number of documentalists increases to 2, and the set of content items is equally divided among them, then the dimensionality of the dataset is $2x|M|x2$, and would be block-diagonal along the user/item dimensions (Figure 6 – (b)). If we had 2 documentalists, each in charge of annotating one Information Element on all items, the tensor dimensionality would still be $2x|M|x2$ but the dataset would be block-diagonal along the user/information element dimension (Figure 6 – (c)).

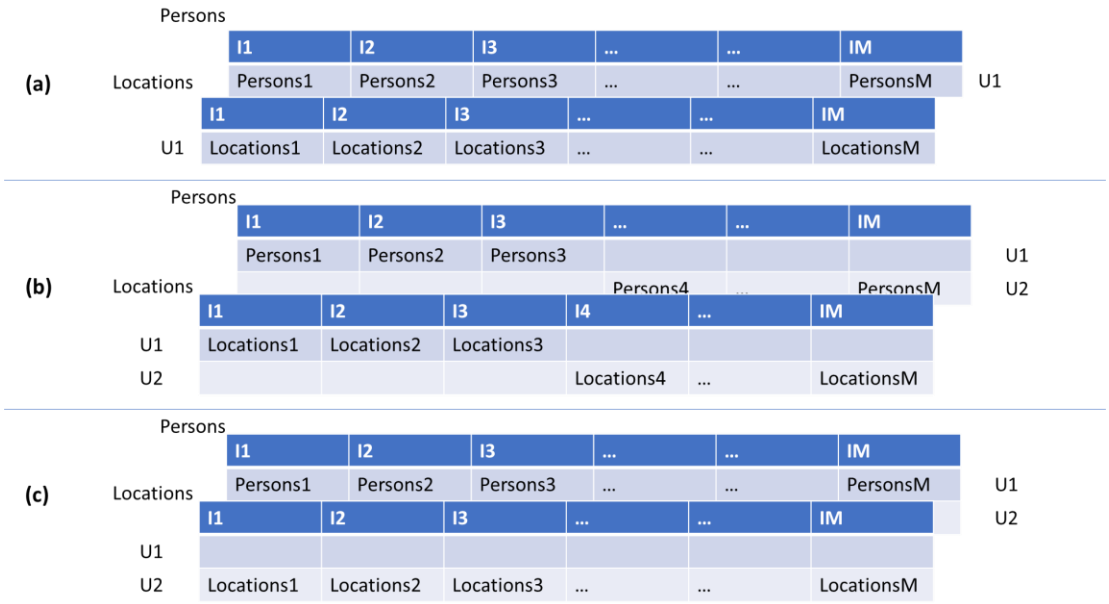


Figure 6. Different dataset tensor-like configurations.

Cross-domain reuse of datasets

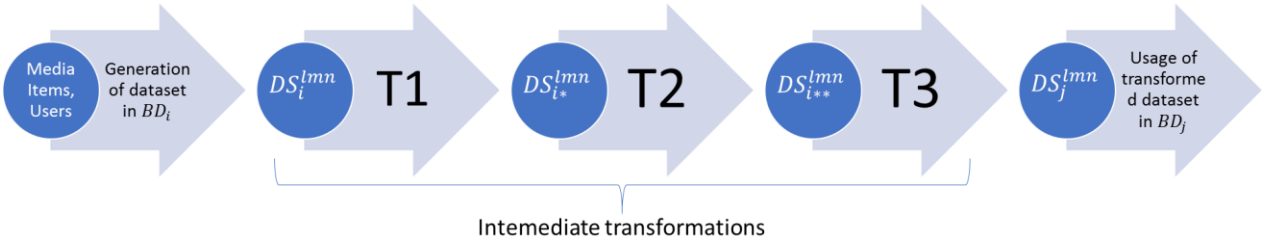
Let us now recall the original context of this dissertation, namely that of objectively characterising the process of reusing archive annotated items as datasets for training AI tools. In the previous section, we introduced a tensor-like representation for datasets and here we use it to further characterise the problem. We can represent the process of cross-domain dataset reuse between business domains BD_i and BD_j as a transformation function T_{ij} between spaces of dataset tensors DS_i^{lmn} and DS_j^{lmn}. Furthermore, we denote with T_{0j} the transformation generating the dataset DS_j^{lmn} in domain BD_j from scratch.

The apparently simple characterisation of the problem is hiding at least three quite complex and articulated issues, e.g.: 1) adaptation/transformation of the source description model to the target one[4]; 2) conciliation of source users' annotations[5] and implementation of user-

[4] E.g., by abstraction, refinement or mapping among information elements.
[5] E.g., by merging or filtering different users' annotations.

dependent annotation of the target domain; 3) addition of new media items. As an example, let us assume the case in which we want to classify OTT media items by genre and we want to reuse a dataset from the archives to train an AI classifier to recognise OTT genres. As we have exemplified in Figure 2, the two domains have different reference taxonomies and different criteria to associate a certain media item to a taxonomy term. Thus, in this case the general adaptation problem is composed of three separate sub-problems: 1) genre taxonomy mapping between source and target[6]; 2) integration of the dataset with new media items from the OTT domain; 3) ground truth media item classification by users of the target domain. This transformation/adaptation process can be partly automatized by e.g. considering static taxonomy mappings, or using more sophisticated approaches employing AI-based techniques. No doubt however, that this process has a non-trivial economic footprint, and as such has to be appropriately designed and budgeted.

A particular family of transformation functions T_{ij} is that for which we can write a complex transformation as a functional composition of intermediate simpler transformations. Intermediate simpler transformations can include atomic operations like for example: a) removal of certain users' annotations; b) merge of users' annotations; c) removal or addition of media items; d) addition or removal of Information Elements; e) addition of users' annotations. Implicitly, intermediate transformations define intermediate Business Domains (adaptation domains), and thus they imply the existence of intermediate classes of Users performing transformations (Figure 7)[7]. A complete formal characterisation of these transformation types is out of scope for this paper, while a detailed discussion about costs related to these transformations will be done in the next Section.



$$DS_j^{lmn} = T3\left(T2\left(T1\left(DS_i^{lmn}\right)\right)\right)$$

Figure 7. Composed dataset transformations.

[6] This mapping is a specific case of a more general case of description model mapping. A description model mapping can be represented as a two-step transformation process: in the first step the target description model is merged with the source description model, in the second step the instances of the target description model information elements are generated from the instances of the source description model's information elements.

[7] Under the assumption that some of these users can be also AI-based machines operating intermediate transformation tasks, a new paradigm comes into the scene: using AI to adapt datasets in order to enable training AI tools across domains. We can call this paradigm AI4AI.

Standard datasets and transfer learning

Naturally, the developed considerations apply unchanged in cases in which the source domains are external to the organisation, e.g., datasets are open and available as part of a scientific initiative or challenge (e.g., (4)). In these cases, the information about the users who created the dataset and their reference annotation rules is normally unavailable. This poses an additional challenge to their reuse in an internal business domain since in principle all ground truth statements should be re-checked beforehand. A somewhat affine concept to the illustrated ones is that of transfer learning (5). Transfer learning is a machine learning technique which uses knowledge acquired in a certain domain or application (e.g., cat breed classification) in the context of another domain (e.g., dog breed classification). The reused knowledge comes normally as a pre-trained AI model which can be fine-tuned in the new domain. The usefulness of this approach is of course limited by the availability of pre-trained models and hence cannot be generalised to all cases of Information Elements.

FORMALISING COSTS CONSIDERATIONS

The previous Sections should have enlightened us to the fact that having annotated assets in the archive is not equivalent to having datasets for AI. In the middle, there are non-trivial adaptation processes to take into account, which bring along actual costs too. In this Section we try to provide a framework to evaluate these costs and assess which is the most appropriate strategy to carry out this adaptation. Additional context to this work is provided in (2) and in (3). The starting consideration is recognising the comparison term against which the adaptation processes is to be referred, namely the one that assumes that the dataset is built from scratch in the target domain.

Let us denote with $\gamma_\epsilon: \boldsymbol{T} \to \mathbb{R}$ a functional associating a cost to a dataset transformation $T_{ij} \in \boldsymbol{T}$, depending on a certain maximum allowed error[8] ϵ. Then we can express the problem of cost comparison of two transformations T_{ij} and T'_{ij} producing the same dataset DS_j^{lmn} from dataset DS_i^{lmn} as:

$$\gamma_\epsilon(T_{ij}) \lesseqgtr \gamma_\epsilon(T'_{ij}) \qquad 1.$$

While in general this is a trivial finding, in practice it can be useful when the process structure of the two transformation is known, for example when they are composed transformations. In this case, we can write:

$$\sum_{i=1}^{N} \gamma_\epsilon(T_{ii+1}) < \sum_{i=1}^{N'} \gamma_\epsilon(T'_{ii+1}) \qquad 2.$$

[8] Costs can heavily depend on the accepted error. We could associate error to measures of information retrieval quality like precision and recall referred between the actual information and the captured ground truth.

Where N and N' are the number of atomic transformations included in the two cases. In this case, the expressed optimisation problem depends on a number of factors, including number, nature and efficiency of each atomic transformation on both sides of the inequality. Out of the many possible concrete configurations in which Eq. 2 can be developed, an interesting case is that in which atomic transformations are performed by automatic tools. In that case, Eq. 2 becomes:

$$\sum_{i=1}^{N} \gamma_\epsilon(T_{ii+1}) < \sum_{i=1}^{N'} \gamma_\epsilon(T'_{ii+1}) + \gamma_\epsilon{}^{CHK}(T'_{ii+1}) \quad 3.$$

Where $\gamma_\epsilon{}^{CHK}(T'_{ii+1})$ represents the cost related do the check-and-correction of errors introduced by the automatic tool performing transformation T'_{ii+1}.

Finally, to impose the condition that adaptation (be it performed manually or AI-supported) is convenient w.r.t. creation from scratch we should write the following condition:

$$\max\left\{\sum_{i=1}^{N} \gamma_\epsilon(T_{ii+1}), \sum_{i=1}^{N'} \gamma_\epsilon(T'_{ii+1}) + \gamma_\epsilon{}^{CHK}(T'_{ii+1})\right\} < \gamma_\epsilon(T_{0j}) \quad 4.$$

Sticking to our running example of genre classification, Table 1 summarises the transformations (column-wise) of three different approaches at the generation of a dataset aimed at training an AI classifier in the OTT domain starting from an archive dataset. Notice that, according to this model, it is assumed that the AI training phase is not able to distinguish, statistically, among the three paths, i.e. that the observed average error ϵ is the same for the three cases. Hence, the quality of the AI model should not be affected by the strategy adopted for dataset creation.

Manual adaptation	***Creation from scratch***	***AI-supported adaptation***
Filtering of source Media Items ($\gamma_{\epsilon,1}$)	Collection of Media Items ($\gamma_{\epsilon,2}$)	Filtering of source Media Items ($\gamma_{\epsilon,1}$)
Manual Genre Taxonomy Mapping ($\gamma_{\epsilon,3}$)	Manual Media Items classification ($\gamma_{\epsilon,4}$)	Automatic Genre Taxonomy Mapping ($\gamma_{\epsilon,5}$)
Collection of additional Media Items ($\gamma_{\epsilon,6}$)		Taxonomy Check&Correction ($\gamma_{\epsilon,7}$)
Manual classification of additional Media Items $\gamma_{\epsilon,8}$		Automatic Media Items Classification ($\gamma_{\epsilon,9}$)
Users' annotation conciliation ($\gamma_{\epsilon,10}$)		Classification Check&Correction ($\gamma_{\epsilon,11}$)
AI training		

Table 1. Comparison of transformations steps in three different cases.

In this example, the winning approach is the one whose sum of costs associated to each traversed row is minimum. Furthermore, Eq. 4 has the following form:

$$\max\{\gamma_{\epsilon,1}+\gamma_{\epsilon,3}+\gamma_{\epsilon,6}+\gamma_{\epsilon,8}+\gamma_{\epsilon,10},\gamma_{\epsilon,1}+\gamma_{\epsilon,5}+\gamma_{\epsilon,7}+\gamma_{\epsilon,9}+\gamma_{\epsilon,11}\} < \gamma_{\epsilon,2}+\gamma_{\epsilon,4}+\gamma_{\epsilon,10} \quad (5.)$$

CONCLUSIONS

With this work, we aimed at contributing in a rigorous way to the discussion about the opportunity of reusing archive assets as ground truth in AI training processes. The default belief is that organisations owning huge archives have a natural advantage in implementing these processes w.r.t. the ones having to start from scratch. To objectify and put under thorough verification this default belief, we presented a formal framework for the representation and analysis of processes related to the reuse of archive assets to train AI tools. We firstly introduced the notion of tensor-like representation of datasets, accounting in a compact manner about the user-dependent nature of annotation and the differences existing in different business domains in terms of description models and annotation rules. We then formalised the problem of dataset reuse across business domains and provided a formal framework for the corresponding costs evaluation. Further future work will consist in refining the theoretical model, and applying it in concrete cases.

REFERENCES

1. A. Messina, “Exploiting Your Data For AI: A Long Road”, EBU Metadata Developers Seminar (MDN), 2019. Geneva, June 2019.
2. A. Messina, “Using Your Archive Metadata For AI Applications”, EBU Production Technology Seminar (PTS), 2020. Geneva, January 2020.
3. A. Messina and D. Airola Gnota, “Automatic Archive Documentation Based on Content Analysis”, International Broacasting Convention (IBC) Conference, 2005. Amsterdam, September 2005.
4. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database”, CVPR09, 2009.
5. L. Pratt and B. Jennings, “A Survey of Transfer Between Connectionist Networks”, Connection Science, 8:2, 163-184, 1996. DOI: 10.1080/095400996116866 .
6. L. Yang , “Artificial intelligence: a survey on evolution, models, applications and future trends”, Journal of Management Analytics, 6:1, 1-29, 2019. DOI: 10.1080/23270012.2019.1570365 .
7. S. M. Chan-Olmsted, “A Review of Artificial Intelligence Adoptions in the Media Industry”, International Journal on Media Management, 21:3-4, 193-215, 2019. DOI: 10.1080/14241277.2019.1695619 .
8. M.I. Jordan, T.M. Mitchell, “Machine learning: Trends, perspectives, and prospects”, Science, 17 JUL 2015 : 255-260.