VisDict: A Visual Dictionary in a Science Gateway

Sandra Gesing
Discovery Partners Institute
University of Illinois
Chicago, IL, USA
sgesing@uillinois.edu

Ewa Deelman
Information Sciences Institute
University of Southern California
Los Angeles, CA, USA
deelman@isi.edu

Rafael Ferreira da Silva
National Center for Computational Sciences
Oak Ridge National Laboratory
Oak Ridge, TN, USA
silvarf@ornl.gov

Michael Hildreth
Physics and Astronomy
University of Notre Dame
Notre Dame, IN, USA
mhildret@nd.edu

Mary Ann McDowell
Biological Sciences
University of Notre Dame
Notre Dame, IN, USA
mcdowell.11@nd.edu

Natalie K. Meyers
Lucy Family Institute
University of Notre Dame
Notre Dame, IN, USA
natalie.meyers@nd.edu

Douglas Thain
Computer Science and Engineering
University of Notre Dame
Notre Dame, IN, USA
dthain@nd.edu

***Abstract*—Effective communication is crucial for the success of academic projects, especially within multidisciplinary teams where researchers come from different backgrounds not only on personal and/or cultural level but also from different disciplines. This can lead to misunderstandings which might not be even obvious in meetings and project plans if the same terms might be used for different concepts. Team members implicitly assume that all parties work with the same definition of terms. The project VisDict addresses the communication between workflow providers and domain researchers via the creation of a visual dictionary in a science gateway so that differences in the perception of terms are easily recognized and can be timely resolved. Dictionaries are used as translation tools between natural languages – the approach for translating from computational science to research domains such as physics and biology is novel. In this paper, we go into detail for our approach to build a dictionary in a science gateway, the lessons learned from carefully curating the first entries, and plans for automating its extension to a large set of relevant terms including their illustrations.**

***Index Terms*—science gateways, visual dictionary, communication, workflows**

I. Introduction

Computational workflows and workflow management systems have proven to be an invaluable asset for quite a few research areas with complex computational methods such as research in the LIGO project [3], genomic analysis, and high-energy physics. Workflow projects are inherently multidisciplinary and to communicate project plans and workflow details between researchers from different disciplines and backgrounds is an error-prone task. The VisDict project addresses effective communication between research domains and workflow providers through the development of a visual dictionary. The dictionary will include the terms and a definition per research domain. The definitions are from a citable resource and/or other well-established dictionary such as the Oxford dictionary for biology [4]. With the notion that "a picture is worth a thousand words", we address that the presentation of concepts in figures not only contributes to the understanding of the term and its definition but that it enriches the experience and makes it easier to grasp the differences between the definitions and concepts between the different domains if applicable for a term.

We started with the strategy to have eleven carefully curated entries for the domains computer science, biology, and physics and use them as a start for the visual dictionary. The manual approach is time consuming and not scalable to fill a dictionary. Thus, we created scripts to automate the selection of terms and suggestions for the visualizations. In this paper, we will go into detail for these steps as well as how we envision to continue to curate the dictionary and the VisDict science gateway that will enable the community to vote on definitions and visualizations.

Presented at Gateways 2022, San Diego, USA, October 18–20, 2022. https://zenodo.org/communities/gateways2022/

II. Background

The challenge of effective communication between domain researchers and computational scientists or research facilitators is recognized by organizers of training for research facilitators. The Virtual Residency for research facilitators [1] offers training for professional skills and especially also training sessions developed by the project CyberAmbassadors [2] to improve the communication between facilitators and domain researchers. The training approach is complementary to VisDict's dictionary – the dictionary can extend the tools available for the training.

Dictionary illustrations "whereby a word is explained by pointing to an object" and the utility of visual dictionaries in cross-disciplinary or multi-language learning inspired this work [5]. In a study by Dziemianko, graphic illustrations in dictionaries had a "statistically significant effect on reception" over unillustrated entries, meaning explanation was significantly less successful (54%) than when color pictures (80%) or line drawings (77%) were present [6].

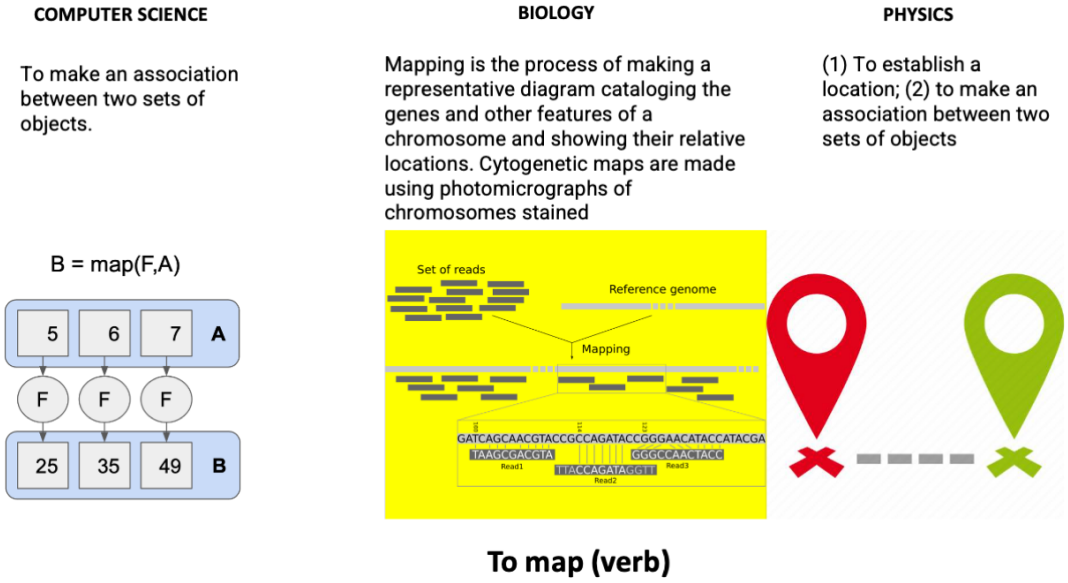

Fig. 1. Example for the Dictionary Entry for "to map" which has three Different Definitions and Illustrations for the three Domains Computer Science, Biology and Physics.

III. Selection of First Terms

In order to select a first set of example terms to test this concept, we selected concepts related to computational workflows whose definitions may be different depending on the domain in which they are referenced. This allows us to demonstrate the variety of perspectives and the potential power of the visual dictionary technique. For example, one of the terms selected was simply "workflow" itself. Even computationally, this word can have a variety of meanings. All of them involve defining some sequence of tasks that can be executed to achieve a desired result. Some workflow patterns can branch or have multiple parallel portions (as in computer science and particle physics), where others are expected to be entirely linear, more like what some would define as a "pipeline", another term that was included. Readers from one field may not be aware that the interpretation of the word is different in another field; exhibiting these difficulties is one of the main purposes of our dictionary.

The eleven terms selected to define the original template for the dictionary were chosen to describe computational workflows or their components. Some describe the workflow itself ("workflow", "pipeline", "sequence"), while others include elements of workflows ("read", "code") and terms related to computation ("scale", "site"). Some of these are especially interesting because they can be both nouns and verbs, also with different meanings. We feel that this first set represents the richness of this vocabulary and poses a good first test of the value of a visual dictionary. Fig. 1 illustrates the dictionary entry for the verb "to map" that shows the different definitions and figures fitting for each domain.

For this first set, the definitions were extracted manually from trusted sources or authored by our group. The images representing the terms were also manually drawn or sourced. This level of curation is not sustainable for a compendium of hundreds of terms. A potential solution to this is discussed in the next section.

IV. Automating Term Search and Visualization Suggestions

To broaden our scope of terms analysis, we have designed and prototyped an open-source toolbox for data capture and analysis of research papers content using Python. The toolbox uses Scrapy [12] and XPath [13] for extracting content from digital libraries. Relevant information (e.g., frequency of words that appear in different sections of the paper, sentiment analysis, etc.) is extracted using machine and statistical learning techniques including NLTK [14], Matplotlib [15], Numpy [16], Pandas [17], Scikit-Learn [18], and GenSim [19]. We are currently working on identifying relevant keywords and on defining a data structure to represent the extracted knowledge. Extracted data will then be formatted and written in a JSON file format and hosted in the project's Science Gateway.

In further steps we will use existing ontologies from different domains, e.g. ontologies for biology [21], to inform keywords and important terms and concepts for VisDict. Some ontologies might have overlap between different domains but most are tailored to their specific use cases. Ontologies can guide the extension to keywords that might be connected to different subdomains in a research domain and provide a wider selection of relevant terms.

We experimented with finding visualizations using various image search APIs, and have shared a Jupyter notebook [11] to search using keywords and download images from different APIs. The sources/APIs used in the notebook for this project

are GoogleImagesSearch by Google Cloud [7], Google Images API by SerpApi [10], `simple_image_download` by PyPI [9] and selenium, chromedriver packages in Python [8]. Google Images API by SerpApi and simple_image_download are the top two sources we used to download the images for this project after comparing the results from the various sources/APIs for our starter set of eleven terms.

We will set up focus groups of students who can vote on visualizations for a set of terms. As start we will set the maximum number of images empirically to 50 and we will explore what number of images creates a well-designed basis set for selections.

V. Voting on Terms and Visualizations

After the first phase of defining dictionary entries via focus groups as described above, we will continue to build content with the citizen science approach. We plan to develop a science gateway based on HUBzero [20] that allows to add terms, their definitions and visualizations via voting for the preferred definitions and illustrations. The selection of terms and the voting steps will be guided by project members and in the long term by a group of dictionary curators to make sure that terms and definitions as well as figures are added that fit to the community and do not reflect any misguiding or offensive material.

It will be crucial to create procedures for creating content and voting that is easy to handle and interesting. Experience with projects on Zooniverse [22], a citizen science platform, shows that people are keen to contribute to science projects — if they feel it is rewarding not only because of them supporting a good cause but is also an enjoyable interaction. We are developing a concept to integrate aspects of gaming such as users can collect points and see a ranking list or combining the voting with rules that users have to overcome some challenges. This is still work in progress and needs more research and evaluation.

VI. Outlook

The next step in the project is setting up the focus group to explore ideal numbers of figures to vote on and to fine-tune our scripts for automatic selection of terms and collections of figures if necessary. The dictionary will always need a human-in-the-loop and community-in-the-loop approach to assure that entries in the dictionary are beneficial and correct. Human-in-the-loop (HITL) often refers to when human labor trains a model, in our case we employ a Community in the loop (CITL) [aka Society in the Loop SITL] strategy for identifying image relevance because we acknowledge role of membership in disciplinary communities as central to our project(physics, genomics, computational science). Disciplinary perspective is crucial to selecting representative illustrations for our terms that have resonance in their respective communities for our visual dictionary strategy to be successful otherwise one discipline's definition (e.g. Computer Science) would overwhelm the others (e.g. physics) which might have fewer numbers of practicioners or image labelers/selectors.

Acknowledgment

We would like to acknowledge support for this work received from NSF award id 2216851 (VisDict). We thank our colleague Ramandeep Makhija, Lucy Family Institute for Data & Society who authored the search API scripts used in this project.

References

[1] http://www.oscer.ou.edu/virtualresidency.php

[2] Astri Briliyanti, Julie Rojewski, T. J. Van Nguyen, Katy Luchini-Colbry, and Dirk Colbry. 2019. The CyberAmbassador Training Program. In Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning) (PEARC '19). Association for Computing Machinery, New York, NY, USA, Article 86, 1–6. https://doi.org/10.1145/3332186.3332218

[3] R. Abbott et al. (LIGO Scientific Collaboration and Virgo Collaboration), "Open data from the first and second observing runs of Advanced LIGO and Advanced Virgo", SoftwareX 13 (2021) 100658.

[4] Hine R, ed. A Dictionary of Biologyverb. 8th ed. Oxford University Press; 2019.

[5] S. I. Landau, Dictionaries: the art and craft of lexicography. New York: Scribner, 1984.

[6] A. Dziemianko . "The usefulness of graphic illustrations in online dictionaries," ReCALL, vol. 34, no. 2, pp. 218–234, May 2022, doi: 10.1017/S0958344021000264

[7] "GoogleImagesSearch - Google Cloud console." https://console.developers.google.com/

[8] "Selenium with Python — Selenium Python Bindings 2 documentation." https://selenium-python.readthedocs.io

[9] J. Dobies, "simple-image-download" https://pypi.org/project/simple-image-download/

[10] SerpApi, "Google Images API". https://serpapi.com/images-results

[11] R. Makhija, "image-search". Lucy Family Institute for Data & Society, 2022. https://github.com/Lucy-Family-Institute/image-search

[12] "Scrapy", 2022. https://scrapy.org/

[13] "XPath", 2022. https://devhints.io/xpath

[14] "NLTK", 2022. https://www.nltk.org/

[15] "Matplotlib", 2022. https://matplotlib.org/

[16] "Numpy", 2022. https://www.numpy.org/

[17] "Pandas", 2022. https://pandas.pydata.org/

[18] "Scikit-Learn", 2022. https://scikit-learn.org/

[19] "GenSim", 2022. https://radimrehurek.com/gensim/

[20] Gesing, S., Stirm, C., Klimeck, G., Zentner, L., Wang, S., Villegas Martinez, B.M., Diaz Eaton, C., Donovan, S., Zhao, L. , Song, C., Kim, I.L., Strachan, A., Zentner, M. and Kalyanam, R. (2022) Open Science via HUBzero: Exploring Five Science Gateways Supporting and Growing their Open Science Communities. Proc. of HICSS-55 (55th Hawaii International Conference on System Sciences), Open Science Practices in Information Systems Research, January 2022, http://hdl.handle.net/10125/79420

[21] Bard JB, Rhee SY. Ontologies in biology: design, applications and future challenges. Nat Rev Genet. 2004 Mar;5(3):213-22. doi: 10.1038/nrg1295. PMID: 14970823.

[22] "Zooniverse", 2022. https://www.zooniverse.org/