Design Principles for Background Knowledge to Enhance Learning in Citizen Science

Kevin Crowston[1](✉), Corey Jackson[2], Isabella Corieri[1], and Carsten Østerlund[1]

[1] Syracuse University, Syracuse, NY 13244, USA
{crowston,ilcorier,costerlu}@syr.edu
[2] University of Wisconsin, Madison, WI 53706, USA
cbjackson2@wisc.edu

Abstract. Citizen scientists make valuable contributions to science but need to learn about the data they are working with to be able to perform more advanced tasks. We present a set of design principles for identifying the kinds of background knowledge that are important to support learning at different stages of engagement, drawn from a study of how free/libre open source software developers are guided to create and use documents. Specifically, we suggest that newcomers require help understanding the purpose, form and content of the documents they engage with, while more advanced developers add understanding of information provenance and the boundaries, relevant participants and work processes. We apply those principles in two separate but related studies. In study 1, we analyze the background knowledge presented to volunteers in the Gravity Spy citizen-science project, mapping the resources to the framework and identifying kinds of knowledge that were not initially provided. In study 2, we use the principles proactively to develop design suggestions for Gravity Spy 2.0, which will involve volunteers in analyzing more diverse sources of data. This new project extends the application of the principles by seeking to use them to support understanding of the relationships between documents, not just the documents individually. We conclude by discussing future work, including a planned evaluation of Gravity Spy 2.0 that will provide a further test of the design principles.

Keywords: Citizen science · Document genre · Boundary objects · Provenance

1 Introduction

The increasing use of automated data-collection instruments has led to an explosion in the amount and diversity of data collected in many settings, from the sciences and medicine to engineering and manufacturing. Making sense of this data deluge requires human perspectives. An increasingly powerful source of human insight at scale is the crowd. A variety of scientific projects currently benefit from engaging volunteers in data analysis—e.g., classifying galaxy shapes

I. Sserwanga et al. (Eds.): iConference 2023, LNCS 13972, pp. 563–580, 2023.
https://doi.org/10.1007/978-3-031-28032-0_43

in the Galaxy Zoo project or identifying exoplanet transits in Planet Hunters—a form of public participation in science referred to as citizen science. Recruiting volunteers to assist with data analysis benefits science from the application of human abilities at a large scale. For instance, Galaxy Zoo data have supported at least 67 publications and Planet Hunters volunteers discovered 120 candidate exoplanets that were not identified by the science team[1]. Volunteers may also benefit by learning about science, provided the opportunity.

Furthermore, we have evidence that with the right support volunteers are capable of more advanced scientific analyses. For instance, Galaxy Zoo volunteers serendipitously discovered a novel kind of galaxy, nicknamed Green Peas [6]. Research on involving volunteers in advanced scientific work suggests that many are both motivated and capable, but need a structured task to be able to contribute [11]. As well, scientific analysis often requires specialized understanding of the nature of the data to effectively navigate and interpret them [13]. Without the proper expertise and knowledge about a dataset and its provenance, volunteers and other less-expert individuals can do little even with large datasets, often being restricted instead to basic analysis.

To address the challenge of enabling crowd members to perform useful and interesting scientific analyses, we aim to develop our understanding of the support they need to collaboratively engage in scientific work. We propose that providing relevant background knowledge will enable even novices to contribute to research. In this paper, we 1) describe the theoretical foundation that guides our search for relevant background knowledge, 2) analyze a citizen-science project to document the ways in which background knowledge is presented to volunteers and 3) use the results of 1 and 2 to develop design ideas about how knowledge should be presented in the follow-on version of the project. The contribution of the paper is to show how the design principles about background knowledge apply in a new setting and how they can be used proactively for design.

2 Theory Development

Past work on citizen science has explored how volunteers learn the task of classifying data. For instance, Jackson et al. [18] found that it benefits volunteer learning and engagement to introduce types of data to be classified gradually rather than all at once. More recent work has shown that as volunteers continue their engagement with a project, the type of learning resource that improves their performance changes: volunteers initially benefit from authoritative resources provided by the science team but later from tools that support their own exploration of the data and interaction with other volunteers [19]. These findings provide a theoretical basis for the current project, but are limited in at least two ways. First, no work has at yet theorized and tested in detail *the nature of the resources* to be provided to support the volunteers. And second, much of the work to date on learning has focused on the basic task of classifying, not the *more advanced work* we seek to support in our project.

[1] https://blog.planethunters.org/2018/11/26/planet-hunters-a-new-beginning/.

To develop principles about the kind of support that will be useful for non-experts to contribute to a project, we draw on research that examines the documents created and used in the process of work [25]. For many collaborators, documents constitute the primary (or even sole) means for knowledge sharing and exchange and form a material instantiation of the work practices. However, to be useful, documents need to be more or less explicit depending on the background knowledge of the intended user [15,20]. Newcomers might need detailed documentation of the work, while an expert can make do with a few bullet points. The latter group holds a shared and practical understanding of the work context that the newcomer lacks. To support the newcomer, a document would have to explicate this knowledge.

To elucidate more precisely the nature of the knowledge needed, we draw on work by Østerlund and Crowston [23], who explored the relationship between free/libre open source software (FLOSS) developers' stock of knowledge and their need for explanations of how to use different documents (e.g., source code, system documentation, project procedures and policies). Participants in FLOSS projects range from core developers with extensive knowledge about the software and software development to peripheral users with limited knowledge. Østerlund and Crowston [23] identified three bodies of theory that speak to the information needs of collaborations that involve such heterogeneous participants: genre theory [3,31,32], boundary objects [29,30] and provenance [14,24]. Each theory addresses the relation between users' stocks of knowledge and their information needs but brings attention to different aspects of the documents that are important.

First, genre theory focuses on the common knowledge people have about documents that they work with. Genre is defined as socially recognized regularities of form and purpose by [32] (e.g., a conference review with a specified form that covers specific topics to inform a publication decision). Members of a relevant community can recognize that a document is of a particular genre, and so know what the expected uses are, but those who do not share that knowledge will need the use, form and expected content spelled out.

Second, the notion of a boundary object addresses how artifacts can bridge between people with few shared points of reference by indicating coincidence boundaries, ideal types or standardized forms [29]. We interpret coincidence boundaries as indicating the value of commonly recognized temporal or participatory boundaries that situate different uses of a document. Ideal types are documents such as diagrams, atlases, or other descriptions that provide an exemplary instance of a document without precisely describing the details of any particular locality, thing, or activity. Finally, standardized forms offers a uniform way to index communicative content and form.

Third, provenance studies speak to how people preserve the history and genealogy of information to alleviate a lack of shared reference points and knowledge that would otherwise impede understanding. For instance, knowing who wrote a document and when can be important to understand its relevance to a current problem.

Combining these three perspectives, Østerlund and Crowston [23] found that documents intended for use by less-knowledgeable members of the community were more likely to be accompanied by explicit statements about:

1. the purpose of the document.
2. the expected form and content of the document. These might even be specified as a standardized form or an ideal type that demarcates specific elements or organization.
3. the context of the document, including the appropriate participants, times and places of the work and the boundaries of the work.
4. the provenance of the document, including the origins of the data and genealogy of its development and use.

In addition to elements suggested by prior theory, Østerlund and Crowston [23] found that documents for novices also expressed a fifth element: the process expectations about the work at hand, that is, what happens to a particular document once it is created.

Of further significance to our project, the study found that FLOSS developers' need for support changed over their engagement with a project. Newcomers required more help understanding the purpose, form and content requirements compared to more advanced participants. As developers gained understanding of the work, they need to understand the boundaries and relevant participants involved in the work (i.e., context), and the information provenance and the process of the information work. This finding suggests directing volunteers to different kinds of background knowledge at different stages rather than simply presenting everything all at once.

Navigating and learning from large scientific datasets comes with unique challenges that differ from learning to contribute to software development processes. FLOSS participants deal with bug reports and source code changes while work with large scientific datasets involves understanding questions like the configuration of instruments and modes of data collection. Nevertheless, generalizing from documents to presentations of data, we believe that providing the identified elements of background knowledge about components of a dataset will support less-expert users in being able to make sense of the data, enabling them to contribute to more advanced analysis.

Based on the review above, we developed the following research questions to address in this paper:

1. What kinds of background knowledge about a dataset are useful for non-experts to be able to understand and work with the data?
2. How does the required knowledge change as volunteers gain experience?
3. What do these findings about background knowledge suggest for the design of future citizen projects?

The first study presented in this paper addresses the first two questions. The second study builds on those results to address the third question.

3 Study 1: Presentation of Background Knowledge

In the first study, we address the first two research questions by carrying out a study of the presentation of background knowledge resources in an existing citizen-science project called Gravity Spy[2] [33].

3.1 Methods

The research uses virtual ethnography [16]. Virtual ethnography adapts traditional ethnographic methods, such as participant observation and in-person interviews, to studying online communities like Gravity Spy. To enhance our understanding of how volunteers in Gravity Spy use background knowledge, we (the authors and other members of the research team) first engaged in Gravity Spy as participant-observers. As participants, we created user accounts, completed requisite training, made classifications, and contributed to project discussions over the course of the first year of the activity, with a lower engagement since then. A first task for all new members of the research team is to go through the same process of initial engagement. We used our position as observers to build knowledge about how volunteers engage with background knowledge on the platform, e.g., what background knowledge resources the system currently provides to volunteers at different stages of engagement and how participants use background information to learn about the project throughout their interaction. We analyzed the resources we identified to determine how they mapped to the categories in the theoretical framework.

We also conducted fifteen interviews, three with members of the Gravity Spy science team and the rest with Gravity Spy volunteers and moderators. Each interview lasted approximately one hour, and was recorded and transcribed. The interviews with scientists focused on how Gravity Spy scientists use data, tools, and other materials to make inferences about relationships between glitches and the auxiliary channels (the task we hope to facilitate in the next version of Gravity Spy). Interviews with volunteers and moderators focused on current background knowledge used to develop insights about the relationships among glitches. Although the inference task is not yet supported, volunteers have attempted to make inferences by linking external materials such as research articles and summary descriptions of detector observation notes on Gravity Spy discussion boards. We also asked moderators questions about how new forms of work (i.e., making inferences) could be supported in a new Gravity Spy interface.

3.2 Setting: Gravity Spy

The Gravity Spy citizen science project [33] incorporated advances in machine learning and new approaches to citizen science to support the Laser Interferometer Gravitational-Wave Observatory (LIGO), a dramatic example of large-scale scientific data collection. LIGO's goal is to detect gravitational waves (GWs),

[2] https://gravityspy.org/.

extremely faint distortions in the fabric of space created by astronomical events such as merging black holes. A challenge for LIGO scientists is that the detectors (one in Hanford, Washington and one in Livingston, Louisiana USA) need to be extremely sensitive to be able to detect GWs, but as a result, they also record orders of magnitude more noise events (referred to as glitches) caused by terrestrial interference or by internal faults or interactions in the detectors. Glitches can obscure or even masquerade as GW signals, so identifying and eliminating their causes is a key activity to improve the detectors [8,12]. These efforts to understand and mitigate these sources of noise, both in the instrument and the data, are collectively referred to as "detector characterization". Gravity Spy supports this work by recruiting volunteers to sort observed glitches into different classes, known or thought to have a common cause. LIGO scientists use the Gravity Spy purified collections to guide their search for the underlying cause of a particular class of glitch, with the goal of eliminating them. We briefly describe the current Gravity Spy project and the volunteers' work to provide context for the discussion of the needed background knowledge.

Classification Work. The Gravity Spy project uses data from the main GW channel from LIGO, a 16 kHz stream of samples [2]. The data-import pipeline extracts two seconds of data around each observed event with a high signal-to-noise ratio, signalling a potential glitch. The data are processed for presentation to humans as spectrograms, specifically, Omega scans [9], a visual representation of the glitch with time on the horizontal axis, frequency on the vertical axis, and intensity of the signal represented by the color from blue to yellow (Fig. 1).

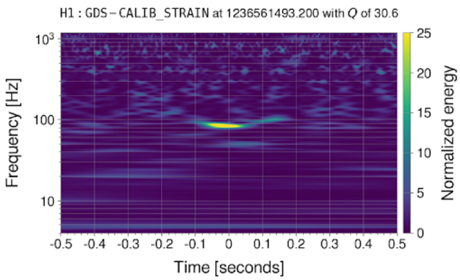


Fig. 1. Spectrogram of a Whistle glitch in the main GW channel

The spectrograms are imported to the Gravity Spy project on the Zooniverse platform [26], where they are presented to volunteers for classification. The classification interface was created using the Zooniverse project builder[3], which enables a Zooniverse project to be created with a few mouse clicks and data uploads. Volunteers label each glitch as being of a known class (23 currently) or "None of the above". To scaffold learning, volunteers progress through

[3] https://www.zooniverse.org/lab.

a series of levels in which they have an increasing number of options for classifying. Newcomers to the project start in level 1 where they given a choice of only Whistle and Blip glitches, which are easy to recognize and distinguish, plus None of the above. Machine learning (ML) supports this process [10]. In initial levels, volunteers are shown only glitches classified by a ML system as being quite likely to be of the classes included in the level. Volunteers also have the option of "None of the above" in case the ML is wrong, meaning that even beginners are doing useful work checking the ML. As volunteers classify and gain experience with glitches, they are promoted to higher levels with more choices, increasing eventually to all 23 classes. At level 5, their attention is focused on glitches that the ML was unable to classify.

Novel Glitch Identification Work. Assigning glitches to the predefined set of glitch classes represents the lion's share of the work done in Gravity Spy. However, some glitches do not fit any known class and so may be examples of as-yet undescribed classes of glitches. If new classes of commonly-occurring glitches were better understood, their causes might be addressed to improve the detectors [28]. Experienced Gravity Spy volunteers identify new classes of glitches by creating and describing collections of "None of the above" glitches with similar novel appearance (collections are a feature of the Zooniverse platform). This work is supported by tools to search for glitches similar to a given glitch and to retrieve metadata for the glitches in a collection. Volunteers can work independently but often collaborate with other volunteers in describing novel classes. Cooperation among volunteers takes place using the Zooniverse platform's Talk forum [17]. Descriptions of suggested new glitch classes are provided to LIGO scientists and if the class is common, volunteers can create a formal proposal that the new glitch be added to the Gravity Spy classification interface. Six new classes have been added to date and many more candidates have been proposed.

3.3 Findings: Background Knowledge in Gravity Spy

In this section, we describe background knowledge resources needed to understand glitches in the LIGO detectors in the current Gravity Spy project, based on our own observation and use of the site. We map these resources to the identified design principles that suggest which will be useful at different stages of engagement with the project. We draw on interviews with active developers to identify background knowledge resources that were not provided by the project developers but that the volunteers identified as helpful.

A Zooniverse project includes multiple venues for presenting background knowledge: the project description and "About this project" pages, a tutorial that is presented to volunteers when they start classifying, a mini-course whose pages are presented interspersed with the classification work, a field guide that can be referenced during the task, a description panel that pops up when a classification is selected, and Talk pages for discussion among volunteers and with the science team. We expect that the About pages and tutorial address the background knowledge needs of newcomers, the mini-course, field guide and detail panels, more experienced volunteers (those over the immediate hurdle of learning how to contribute), and the Talk pages, advanced volunteers. This progression shows a transition from authoritative to collaborative resources [19].

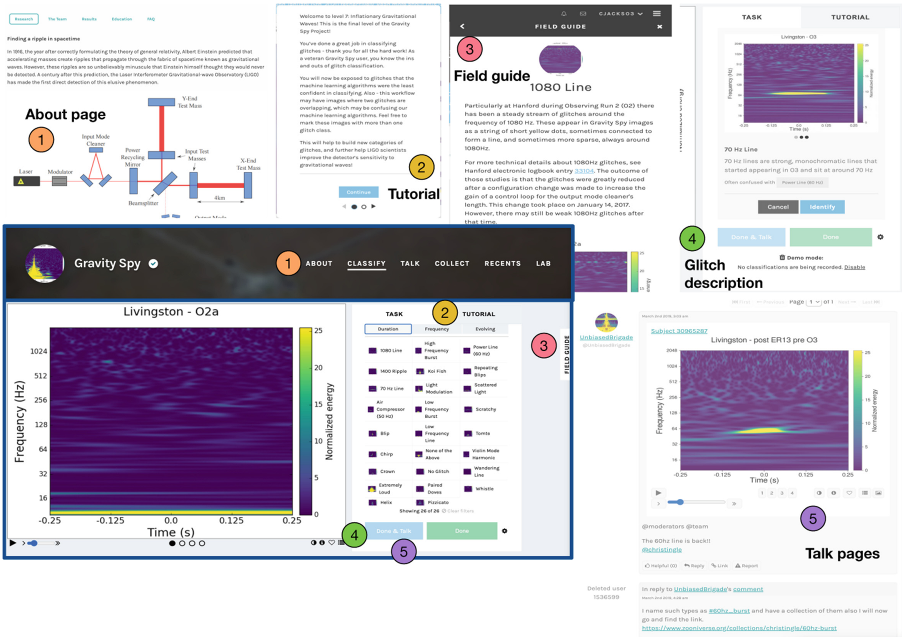


Fig. 2. The Gravity Spy classification interface is on the lower left, with the spectogram of the glitch to be classified on the left and the possible classes on the right. The numbered circles indicate the background knowledge resources provided, with examples above and to the right.

For the classification task in Gravity Spy, the documents that volunteers need to understand are the spectograms that they classify (shown on the left side at the lower left in Fig. 2). In the current Gravity Spy, the "About" pages (1 in Fig. 2) present the goals of LIGO, how the detectors work, what glitches are, the goals of the Gravity Spy project and the research team. The pages also provide links to published papers about Gravity Spy and to other reading about LIGO and the detector. Each level has its own tutorial (2 in Fig. 2) to introduce features added at that level. The tutorial is automatically shown the first time a volunteer starts a level and is available afterwards on demand. The level 1 tutorial, shown to newcomers to the project, explains what a spectogram is, how to perform a classification using the Zooniverse interface and how volunteers are promoted to advanced levels. The mini-course (not shown) presents information about LIGO, as much to keep volunteers' interest than because of its immediate relevance to the task. The field guide (3 in Fig. 2) describes each of the 23 known classes of glitch with examples of their appearance, as do the popups that appear when a class is selected for a glitch (4 in Fig. 2). Finally, the Talk pages (5 in Fig. 2) includes boards to chat, ask for help, report bugs, comment on specific glitches or to discuss the science behind the project. Additional boards were created later to discuss and propose potential new glitch classes discovered by the volunteers. Some experienced volunteers act as moderators for the Talk pages and often answer questions from other volunteers.

The design principles developed above suggest that newcomers require help understanding the purpose of documents and their form and content. Reflecting this ordering, the current Gravity Spy About pages and tutorial describe the purpose of a spectogram, i.e., to show a glitch in a human readable format, its form and what content it contains, namely a glitch. More established users need to understand the boundaries and relevant participants involved in the work, that is, how the work they are doing connects with other tasks and other participants. The Gravity Spy project initially did not provide this information. However, advanced Gravity Spy volunteers have posted a range of potentially useful information to the Talk pages, an example of collaboratively-created background knowledge resources [19]. These include discussions of how the spectograms are created and links to LIGO aLogs[4], which record work done on the detectors, linking the work of the LIGO scientists to the work of the volunteers.

As noted above, the advanced work in Gravity Spy consists of collecting examples of potential new glitch classes and describing some of these classes in a glitch proposal document. This work introduces two new kinds of documents that must be understood, specifically collections and glitch class proposals. Volunteers often collaborate to create these documents. Gravity Spy at present does not explicitly describe this work nor provide relevant background knowledge beyond the knowledge needed to do the initial classification task. Again, the volunteers have created Talk posts that explicate the process. The project scientists did create a template for a glitch class proposal, consistent with Østerlund and Crowston [23]'s finding that such standard forms are used to regulate communicate between groups with different levels of background knowledge, in this case volunteers and science team members. Accepted glitch class proposals also constitute a kind of ideal type for creating new proposals. In summary, the framework seems to capture the kinds of background knowledge provided in Gravity Spy as well as identifying lacunae (RQ1), and how these resources change as volunteers gain experience (RQ2).

4 Study 2: Theory-Driven System Design

In this section, we present the second study, which seeks to use the design principles developed above to proactively guide the design of a system to address a novel problem (RQ3). We describe the novel problem, how that problem is handled by experts and the suggestions from the principles about how to present necessary background knowledge to enable volunteers to take on the task.

[4] https://alog.ligo-la.caltech.edu/aLOG/.

As noted above, a finding of our study is that the current Gravity Spy system does not provide authoritative resources to support the advanced work of identifying new glitch classes but that volunteers have created some. Still, volunteers face challenges identifying and describing new glitch classes in a useful way. The hope is that glitches in a new class have a common cause that can be addressed. However, at present volunteers have limited knowledge about the underlying mechanisms within the detectors that generate glitches, nor can they explore those mechanisms. As a result, new glitch class identification is done phenomenologically, i.e., by grouping glitches with similar appearance (witness the fact that volunteer-identified glitch classes are named by shape, e.g., Helix or Crown, in contrast to most LIGO-identified classes that are named by cause, e.g., Whistle or Scattered Light). This approach has been effective in identifying new glitch classes. However, the essential next step of identifying causes requires the attention of the overloaded LIGO science team. In this section, we describe how we are using what we have learned about background knowledge to design a new citizen-science project that will enable volunteers to take on some of this analysis work, addressing our third research question.

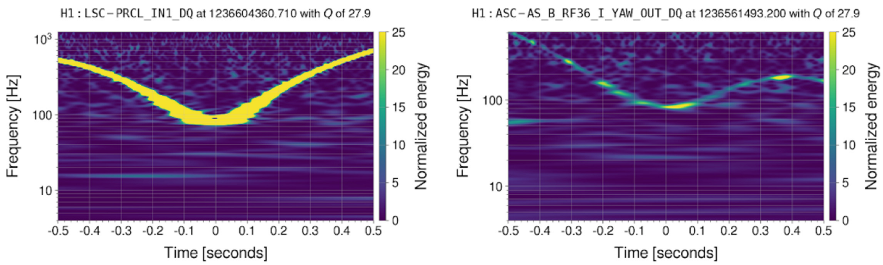
4.1 Methods

To identify what resources would be useful to support this task, we carried out interviews with experts as described above for Study 1. These interviews gave us an understanding of the task to be supported and background knowledge resources that might be useful. The resources identified by the experts were sorted by the categories in the theory and to modes of delivery in the project.

4.2 Data-Centred Approaches to Glitch Analysis

We start by describing how professional LIGO scientists address the task. To explore the cause of glitches (i.e., what is happening in the detector or the environment that causes particular classes of glitches), LIGO scientists carry out studies using what are called auxiliary channel (AC) data. Along with GWs, the LIGO detectors record more than 200,000 channels of data per detector from a diverse set of sensors that continuously measure every aspect of the detectors and their environment (e.g., equipment functioning, activation of components, seismic activity or weather) [21,22]. This dataset holds clues to the cause of glitches, but the large volume of data demands ways to transform this massive volume of data from disparate sources into useful information.

Currently LIGO uses a number of algorithms (e.g., hVeto [27], iDQ [4], Karoo GP [8]) that identify statistically-significant correlations between a loud event occurring in the main GW channel (a likely glitch) and an event in one of the other channels. Since different classes of glitches are created by different mechanisms, they are correlated with diverse ACs. As useful as these tools are for providing clues to the causes of glitches, statistical correlations represent an incomplete picture and do not clearly point to causality. Some channels experience loud events frequently, so the fact that they correlate with a glitch might not be informative. Channels have complicated interdependencies (e.g., because of being in the same location or actually dependent on each other through feedback loops), so many channels can show correlation with the same glitch. As a result, a channel may be a statistically-significant witness for a class of glitch even though it is not actually close to the root cause. A further issue is that only some of the mechanisms connecting parts of the detector are well understood. Mechanisms can be complex and non-linear, may involve complicated interactions (e.g., between environmental conditions and detector functioning) and some are yet to be discovered. Much work is needed to determine if highly-correlated events in the ACs point to the root cause of the glitch.



(a) (b)

Fig. 3. Spectrograms of two auxiliary channels related to the glitch in Fig. 1, (a) power recycling cavity length (PRCL) and (cb alignment control channel (ASC).

Figure 3 illustrates the exploration process as currently performed by LIGO scientists. Simply looking at a spectrogram of a glitch from the main GW channel (Fig. 1 above) does not show a very obvious morphology. The slight change in frequency hints at the type of glitch, a Whistle, but information from auxiliary channels is needed to understand its cause. A first step in the exploration is a closer comparison of the morphology of the glitch and correlated channels. For our Whistle, looking at the power recycling cavity length (PRCL) channel (Fig. 3a) one finds an event that looks like a louder version of the Whistle glitch; the same shape is present in the GW channel (Fig. 1), but at a much lower amplitude, largely obscured by noise. Other channels may show similar patterns-e.g., Fig. 3b, an alignment control channel (ASC)-but not as strongly as PRCL. Understanding the layout of the detectors and the provenance of the dataset helps to make sense of the root cause of the Whistle: the GW and the PRCL channel (among others) witness radio frequencies; different radio frequency oscillators move closer and farther apart in frequency, creating a varying beat note that is the Whistles' unique pattern.

Whistles provide a particularly clear example of a connection between glitches and events in other channels. More challenging classes of glitches require exploring correlations between multiple manifestations of the class and relevant ACs over longer periods of time to develop a full picture. Looking through spectrograms comparable to those in Figs. 1 and 3 (but for hundreds of ACs) over many different glitches can provide hints to the root cause of the glitch, as the same pattern of channels reappear in association with the same kinds of glitches. However, interpreting these patterns requires understanding likely mechanisms of glitch creation.

4.3 Enabling Volunteers to Engage in Glitch Analysis

At the moment, the analysis of novel glitches described above is done only by the LIGO scientists, and their analyses are limited by the time they have available. Based on our understanding of what is needed to enable non-experts to explore complex datasets, developed in the study described above, we believe that we can enable citizen scientists to carry out some of the time-consuming analysis required for the novel classes of glitches that they are already involved in identifying. To do this, we will provide volunteers with access to auxiliary channel data and, more importantly, support them in learning about the detector and the data it records, e.g., by providing relevant background information about the channels and the process by which channels influence each other. Developing and evaluating this system will serve as a further test of the theoretical framework articulated above.

Specifically, our plan is to develop a new citizen-science project, Gravity Spy 2.0. Volunteers will move through different tasks as they contribute to the analysis and build their knowledge, as shown in Fig. 4. In the first task, knowledge will be built while examining individual glitches and vetting their relation to activity in the various ACs. We have identified a subset of several hundred channels that are most informative to use in the project. This task will be performed in a Zooniverse project-builder project. As in the current Gravity Spy system, we plan to introduce glitches and ACs gradually so volunteers have time to learn the nature of that set of glitches or channels. This staging will be supported by

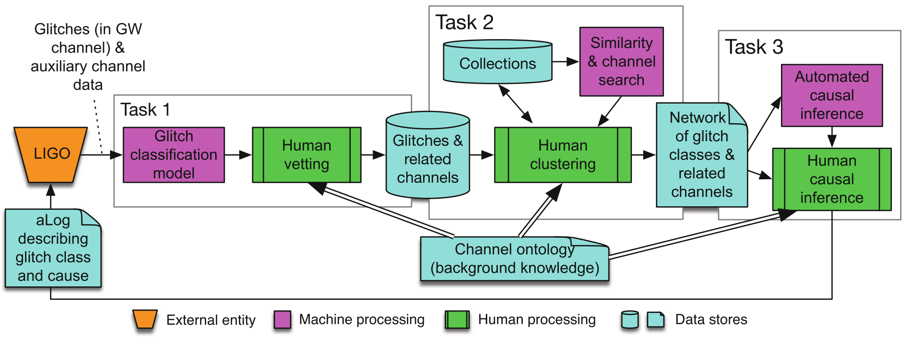


Fig. 4. Flow of data through volunteer and ML processing for Gravity Spy 2.0.

doing an initial sort of glitches using the ML glitch classification models created for Gravity Spy.

In the second task, volunteers will examine collections of glitches and the ACs identified as related in the first stage to identify recurring patterns of connections for a particular class of glitch, and ultimately (in the third task) to deduce which ACs are the causes of those glitches. Both tasks will be supported by additional ML processing, to search for glitches with similar appearance and pattern of related ACs or to draw causal inferences from the connections.

4.4 Background Knowledge to Support Gravity Spy 2.0

In this section, we present our ideas for designing background knowledge for Gravity Spy 2.0, considering primarily Task 1, our current design focus. As with the original Gravity Spy, we expect newcomers to first need to understand the form and content of the documents, through material presented in the About pages and tutorial. Much of the background knowledge material developed for Gravity Spy is still applicable. Indeed, it would likely be beneficial for volunteers to have experience with Gravity Spy 1.0 before engaging with 2.0. However, for Task 1, the materials will also need to explain how the spectograms present information from different ACs and what those are.

The design principles suggest that more established volunteers need help understanding information provenance and the process of the work. Provenance information for LIGO AC data includes what kind of detector collected them (e.g., a seismometer vs. a magnetometer), which is necessary for understanding their implications for glitch formation. To understand provenance, a basic understanding of the parts of the detector will be necessary. One resource is published descriptions of the detector and its subsystems, e.g., [1,5], along with papers describing glitches and how they are characterized, e.g., [2,12,21,22]. These papers might be linked directly or summarized. A list of acronyms[5] will also be helpful for decoding the detector descriptions and the channel names.

A key element in the system will be an ontology of the ACs that presents the background knowledge needed to understand each channel. We are currently building an initial ontology from existing LIGO documentation, with input from LIGO experts. For instance, LIGO maintains a public website[6] that describes the physical and environmental monitoring sensors and a private website describing the instrumental channels. The ontology will be refined throughout our project. A limitation of the Zooniverse project builder is that the field guide is a simple list, making it unusable for presenting information about hundreds of channels. To get around this limitation, we will present the information on a Wiki. Part of the Wiki page will be populated from structured data about each channel (see Fig. 5 for a prototype of a channel page). The Wiki will also allow description and exploration of clusters of related channels, e.g., those in the same subsystem or at the same physical location in the detector (the links in the breakdown of

[5] https://dcc.ligo.org/M080375/public.
[6] http://pem.ligo.org/.

the component name). A key benefit of presenting the information on a Wiki is that volunteers will be able to add to it, thus supporting individual exploration and collaborative background knowledge creation in a structured way.

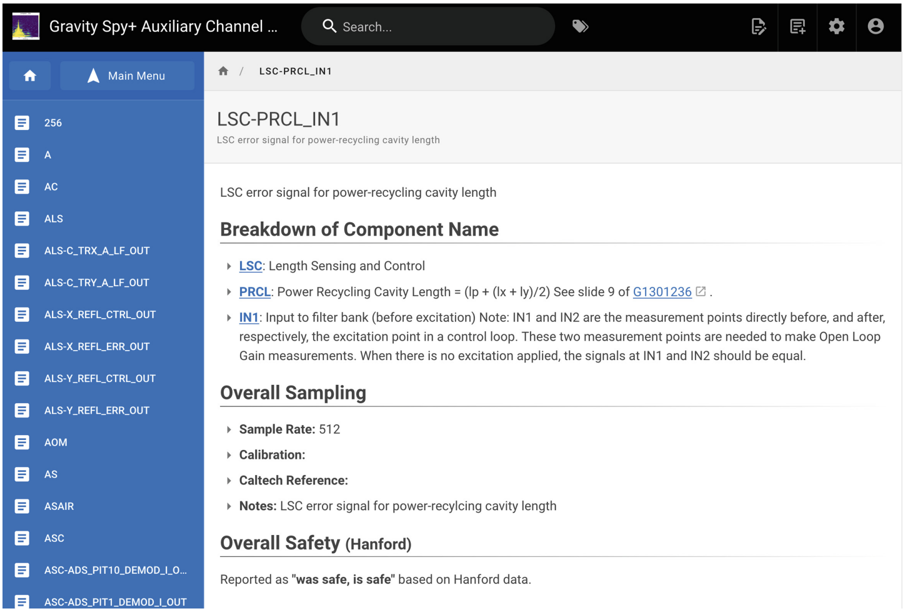


Fig. 5. Prototype Wiki page for the channel shown in Fig. 3a.

4.5 Background Knowledge for Understanding Related Documents

The design principles presented above describe the kinds of background knowledge need to understand single documents. A needed extension to the model is that, in addition to understanding documents individually, we also want the volunteers to understand possible relations between documents, i.e., how a glitch recorded in the GW channel relates to a signal in one of the ACs. We believe that the design principles developed above also apply to describing the background knowledge needed to understand these relations.

Specifically, we believe that newcomers to the task will first need to understand the purpose of the relations and their form and content. The purpose will be described in the About pages, namely, to identify which channels may be part of the processes creating glitches. Form and content in this case refers to how the two spectograms are related. As noted above, glitches do not simply appear in the same form in different ACs, so volunteers will need to learn the form of the relation (i.e., what a Whistle glitch looks like in other channels). We plan to add information about relationships to the AC ontology, e.g., information showing the form of a relationship between channels as it is discovered. As it would be impractical to capture all combinations of hundreds of channels, we will focus on describing how a signal in the AC affects the GW channel for a particular kind of glitch. For instance, the page shown in Fig. 5 could describe for which glitch classes it has been observed to be active; a page for a particular glitch class, the seemingly related channels, perhaps with examples. However, it could be that some other combinations are interesting and worth describing, e.g., channels that seem to be frequently active in combination. An advantage of presenting the information on a Wiki is that the volunteers will be able to extend the channel information as they discover interesting relations.

A major complication is that we lack training material for most combinations of glitches and ACs. To fill this lacuna, we will use the volunteers' contributions to identify relations. To do so, we need to develop a way of describing the relations between spectograms in a few basic terms (e.g., "identical", "same shape, reduced intensity", "truncated", "no relation") that volunteers can reliably identify. In task 2, volunteers will examine collections of glitches to identify which ACs are reliably related and in what way. This identification can then feedback to support the volunteers working on task 1 and eventually to train ML systems.

Once volunteers are past the initial hurdle of learning how to interpret the form and content of related channels, we expect that they will need information about provenance and process. In this setting, provenance means understanding the origin of the relation between the GW and ACs, that is, what about the detector causes those channels to interact? The plan is to provide a description of the detector functioning that should support volunteers in understanding these connections. Finally, to support the most advanced users, we need to present information about the context of the work, specifically, how identifying relationships will support further work with this dataset. Such information can be presented in the Talk discussions or added to the Wiki.

5 Conclusion

We are currently building the system described above. The Zooniverse project builder makes it straightforward to present spectograms to volunteers and to collect their judgements about the relationships. The difficult part in building the initial phase of the project is determining what kinds of background knowledge volunteers need to make sense of the images being presented and to understand whether there is or is not a relationship. By drawing design principles from the theories presented above, we have developed a starting set of ideas about what kinds of background knowledge are important and are now developing materials to populate the site (e.g., the Wiki pages shown in Fig. 5). We are also investigating the contribution of ML processing, e.g., to pick glitches of a particular class to show a volunteer or to cluster channels with a similar relationship to a glitch.

We are currently conducting focus groups with advanced volunteers to refine the design. Participants have suggested additional resources that they have found helpful that we are including in the design. For instance, one volunteer pointed us to a Ph.D. thesis [7] describing the control system for the Virgo detector, which operates on a similar principle to LIGO. A few volunteers will be interested in such resources, while more may benefit from excerpts or summaries on the Wiki.

In future studies, we will evaluate the usefulness of identified elements of background knowledge by analyzing system log data which contains information about elements that volunteers interact with. Through this analysis, we hope to uncover which knowledge (e.g., form and content, purpose, etc.) about glitches and ACs is important in supporting less-expert users in being able to make sense of the data. We will also evaluate learning enhancements by correlating use of background knowledge with volunteers' performance and engagement. Since we expect background knowledge will enhance learning, we can identify whether volunteers who used certain resources produced more advanced analysis. We can also test whether the framework applies to understanding document relationships, as well as documents individually.

Overall, we expect our ongoing research to provide useful and novel insights about the kinds of background knowledge that are effective in enhancing the abilities of non-experts to conduct advanced data analysis. We expect our results to be informative in the many settings where less expert users want to be able to contribute to a complex on-going project. The design principles articulated in Sect. 2 describe the kinds of background knowledge that should be supplied and how these should be ordered. For instance, in a biodiversity project like Snapshot Serengeti[7], we expect newcomers to benefit from explanations of the purpose, form and content of the documents they will encounter (e.g., the photographs and descriptions of the species). Information about the context of the work or provenance of the images might be deferred until those elements are mastered. Our experience in building and operating Gravity Spy 2.0 will provide a needed test and perhaps update of these principles. Armed with these results, future project developers will be better able to scaffold the introduction of relevant background knowledge to smooth volunteers' entry into and progression through their projects.

Acknowledgments. Partially funded by grants from the US National Science Foundation, INSPIRE 15–47880 and HCC 21–06865. The authors thank the Gravity Spy volunteers who participated in this research. Without their contribution, the project would not exist.

References

1. Aasi, J., et al.: Advanced LIGO. Class. Quantum Gravity **32**(7), 074001 (2015)
2. Abbott, B.P., et al.: A guide to LIGO-Virgo detector noise and extraction of transient gravitational-wave signals. Class. Quantum Gravity **37**(5), 055002 (2020)
3. Bazerman, C.: Systems of genres and the enactment of social intentions. In: Freedman, A., Medway, P. (eds.) Genre and the New Rhetoric, pp. 79–101. Taylor and Francis (1995)

[7] https://www.zooniverse.org/projects/zooniverse/snapshot-serengeti.

4. Biswas, R., et al.: Application of machine learning algorithms to the study of noise artifacts in gravitational-wave data. Phys. Rev. D **88**(6), 062003 (2013). arXiv: 1303.6984
5. Buikema, A., et al.: Sensitivity and performance of the advanced LIGO detectors in the third observing run. Phys. Rev. D **102**(6), September 2020
6. Cardamone, C., et al.: Galaxy zoo green peas: discovery of a class of compact extremely star-forming galaxies. Mon. Not. R. Astron. Soc. **399**(3), 1191–1205 (2009)
7. Casanueva Diaz, J.: Control of the gravitational wave interferometric detector Advanced Virgo. Ph.D. thesis, Université Paris-Saclay (ComUE) (2017)
8. Cavaglia, M., Staats, K., Gill, T.: Finding the origin of noise transients in LIGO data with machine learning. Commun. Comput. Phys. **25**(4) (2019). arXiv: 1812.05225
9. Chatterji, S., Blackburn, L., Martin, G., Katsavounidis, E.: Multiresolution techniques for the detection of gravitational-wave bursts. Class. Quantum Gravity **21**(20), S1809–S1818 (2004)
10. Coughlin, S., et al.: Classifying the unknown: discovering novel gravitational-wave detector glitches using similarity learning. Phys. Rev. D **99**(8), 082002 (2019)
11. Crowston, K., Mitchell, E., Østerlund, C.: Coordinating advanced crowd work: Extending citizen science. Citizen Science: Theory and Practice 4(1) (2019)
12. Davis, D., et al.: LIGO detector characterization in the second and third observing runs. Class. Quantum Gravity **38**(13), 135014 (2021)
13. Finholt, T.A., Olson, G.M.: From laboratories to collaboratories: a new organizational form for scientific collaboration. Psychol. Sci. **8**(1), 28–36 (1997)
14. Gilliland-Swetland, A.: Electronic records management. Ann. Rev. Inf. Sci. Technol. **39**(1), 219–253 (2005)
15. Harper, R.: Inside the IMF. Routledge (2009)
16. Hine, C.: Virtual Ethnography. SAGE Publications Ltd, SAGE Publications Ltd, Apr 2000
17. Jackson, C., Crowston, K., Østerlund, C., Harandi, M.: Folksonomies to support coordination and coordination of folksonomies. Comput. Supported Cooperative Work (CSCW) **27**(3–6), 647–678 (2018)
18. Jackson, C.B., et al.: Teaching citizen scientists to categorize glitches using machine-learning-guided training. Computers in Human Behavior 105 (2020)
19. Jackson, C.B., Østerlund, C., Harandi, M., Crowston, K., Trouille, L.: Shifting forms of engagement: Volunteer learning in online citizen science. In: Proceedings of the ACM on Human-Computer Interaction 4(CWCW), 36 (2020)
20. Latour, B.: Pandora's Hope: Essays on the Reality of Science Studies. Harvard University Press (1999)
21. Nguyen, P., Schofield, R.M.S., Effler, A., Austin, C., Adya, V., Ball, M., Banagiri, S., Banowetz, K., Billman, C., Blair, C.D., et al.: Environmental noise in advanced LIGO detectors. Class. Quantum Gravity **38**(14), 145001 (2021)
22. Nuttall, L.K.: Characterizing transient noise in the LIGO detectors. Philosophical Trans. Roy. Soc. A Math. Phys. Eng. Sci. **376**(2120), 20170286 (2018)
23. Østerlund, C., Crowston, K.: Documentation and access to knowledge in online communities: know your audience and write appropriately? J. Am. Soc. Inform. Sci. Technol. **70**, 619–633 (2019)
24. Ram, S., Liu, J.: A semantic foundation for provenance management. J. Data Semant. **1**(1), 11–17 (2012)
25. Shankar, K., Hakken, D., Østerlund, C.: Rethinking documents. In: The Handbook of Science and Technology Studies, 4 edn., pp. 59–86. MIT Press, Cambridge (2017)

26. Simpson, R., Page, K.R., De Roure, D.: Zooniverse: observing the world's largest citizen science platform. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 1049–1054. ACM (2014)
27. Smith, J.R., et al.: A hierarchical method for vetoing noise transients in gravitational-wave detectors. Class. Quantum Gravity **28**(23), 235005 (2011). arXiv: 1107.2948
28. Soni, S., et al.: Discovering features in gravitational-wave data through detector characterization, citizen science and machine learning. Class. Quantum Gravity **38**(19), 195016 (2021)
29. Star, S.L.: The structure of ill-structured solutions: boundary objects and heterogeneous distributed problem solving. In: Gasser, L., Huhns, M.N. (eds.) Distributed Artificial Intelligence, vol. 2, p. 37–54. Morgan Kaufmann (1989)
30. Star, S.L., Griesemer, J.R.: Institutional ecology, 'translations' and boundary objects: amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. In: Social Studies of Science, vol. 19, pp. 387–420. Sage (1989)
31. Swales, J.M.: Genre Analysis: English in Academic and Research Settings. Cambridge University Press (1990)
32. Yates, J., Orlikowski, W.J.: Genres of organizational communication: a structurational approach to studying communications and media. Acad. Manag. Rev. **17**(2), 299–326 (1992)
33. Zevin, M., et al.: Gravity spy: integrating advanced LIGO detector characterization, machine learning, and citizen science. Class. Quantum Gravity **34**(6), 064003 (2017)