Annotating PDF documents for the GeoArchive

Sky Bristol[1]
[1]USGS

Aug 19, 2021
In Development
Share
This protocol is published without a DOI.
GeoArchive

Sky Bristol

DISCLAIMER

DISCLAIMER – FOR INFORMATIONAL PURPOSES ONLY; USE AT YOUR OWN RISK

The protocol content here is for informational purposes only and does not constitute legal, medical, clinical, or safety advice, or otherwise; content added to protocols.io is not peer reviewed and may not have undergone a formal approval of any kind. Information presented in this protocol should not substitute for independent professional judgment, advice, diagnosis, or treatment. Any action you take or refrain from taking using or relying upon the information presented here is strictly at your own risk. You agree that neither the Company nor any of the authors, contributors, administrators, or anyone else associated with protocols.io, can be held responsible for your use of the information contained in or linked to this protocol or any of our Sites/Apps and Services.

ABSTRACT

This protocol provides steps to catalog NI 43-101 Technical Reports within the Zotero reference management system and use PDF highlighting and annotation to identify specific citation metadata and other descriptive information from the texts. It is part of a body of work being developed to manage a GeoArchive digital library of important reference materials in minerals assessments, energy assessments, and other work of the US Geological Survey.

PROTOCOL CITATION

Sky Bristol 2021. Annotating PDF documents for the GeoArchive. **protocols.io** https://protocols.io/view/annotating-pdf-documents-for-the-geoarchive-bxhspj6e

CREATED

Aug 19, 2021

LAST MODIFIED

Aug 19, 2021

PROTOCOL INTEGER ID

52498

DISCLAIMER:

DISCLAIMER – FOR INFORMATIONAL PURPOSES ONLY; USE AT YOUR OWN RISK

The protocol content here is for informational purposes only and does not constitute legal, medical, clinical, or safety advice, or otherwise; content added to protocols.io is not peer reviewed and may not have undergone a formal approval of any kind. Information presented in this protocol should not substitute for independent professional

judgment, advice, diagnosis, or treatment. Any action you take or refrain from taking using or relying upon the information presented here is strictly at your own risk. You agree that neither the Company nor any of the authors, contributors, administrators, or anyone else associated with protocols.io, can be held responsible for your use of the information contained in or linked to this protocol or any of our Sites/Apps and Services.

BEFORE STARTING

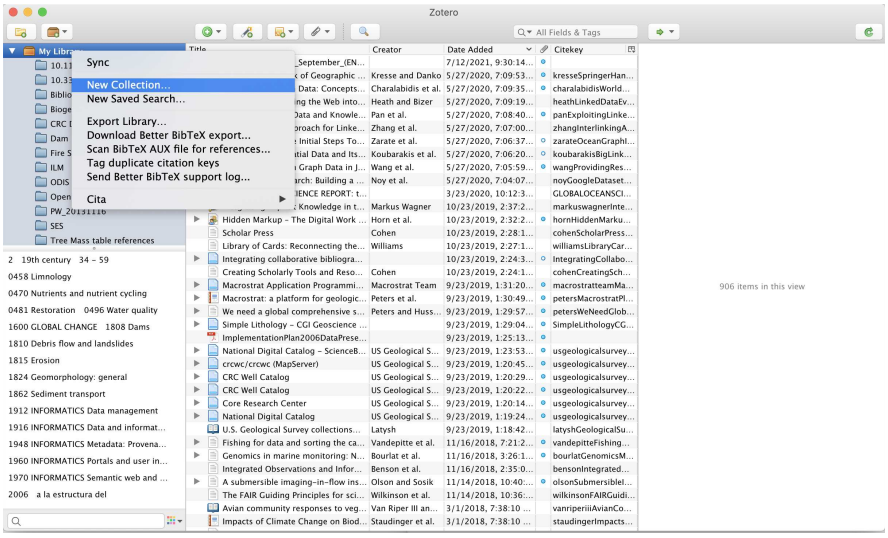
Determine what PDF annotation tool you are going to use. You likely have something built in by default for your operating system to open and read PDFs. On Mac this may be the Preview app or something else. On Windows, this could be a free version of Adobe Acrobat Reader.

Setup Collection in Zotero

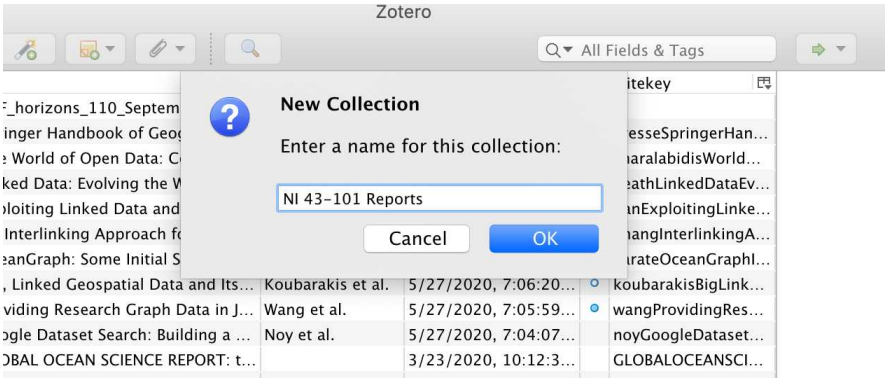
1 Setup a local collection for the reports/documents you want to start annotating in Zotero.

We will eventually move this to a shared library or move the reports out of Zotero into some other tool. In the near term, this provides a way to quickly get files organized into a new space, annotate them, and start working within a reference management framework. This process will copy the files from either some other place on your local hard drive or the network file share.

1.1 Create collection



1.2

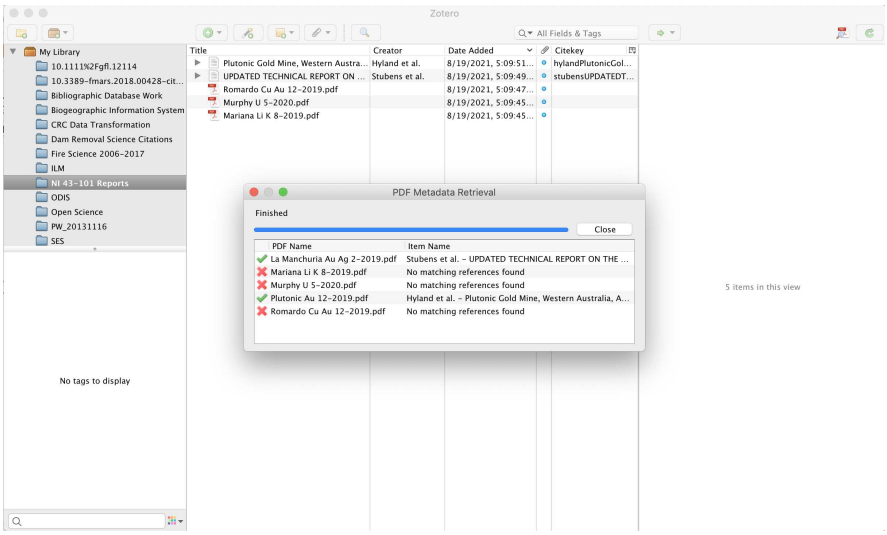


You might call this NI 43-101 Reports or some other name of your choosing. This is currently your local working space for these files, so it doesn't particularly matter what you call it.

Import Files to Zotero

2 Copying the files into Zotero means that Zotero will place them into its own directory structure. This facilitates Zotero working with the files, linking to them with internal references, and other actions. For our purposes, this simply makes sure that we have all of the files in one place along with any metadata we can start developing for them so that we can move it all to a shared, group library.

2.1 Drag and drop files



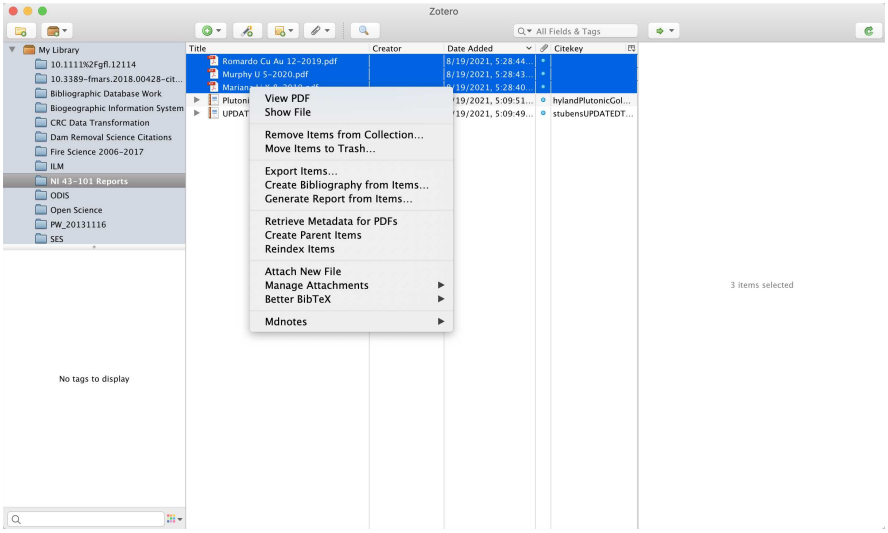
Dragging and dropping PDF files to a Zotero collection will automatically run a process to try and extract information from the PDFs to fill out citation metadata. This is hit and miss for many PDF files like the NI 43-101 reports. Some of them are well structured with at least some metadata that Zotero will interpret.

Click Close once this process is complete.

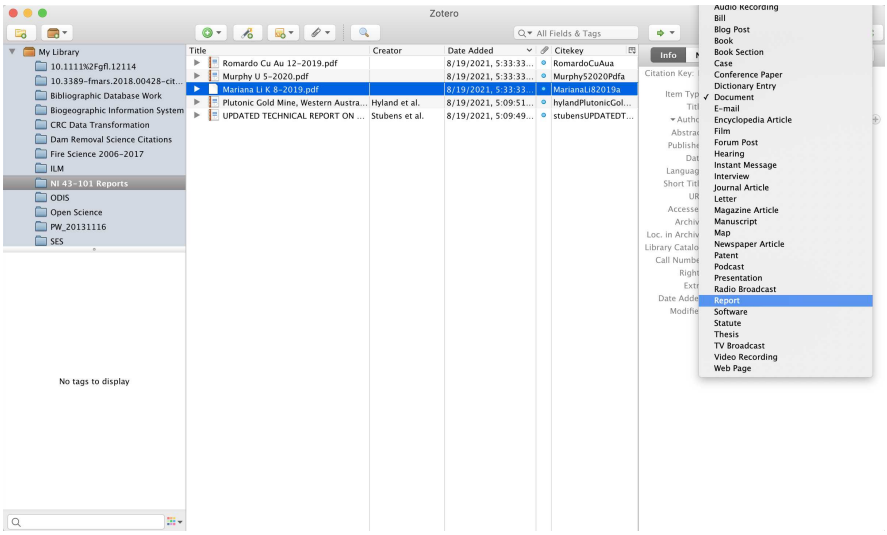
Establish Proper Item Types

3 After dragging and dropping a batch of files into Zotero, you may see that some have been brought in as journal article type items while others are simply files that do not yet have a Zotero item created for them. Zotero defaults to "thinking" that everything is a journal article as this is the most common use case. We need to establish items for everything and set everything to the appropriate type. This section covers that work.

3.1 Select files that were not processed to create items, right click, and select Create Parent Items from the list of options.



3.2 Change all item types to "Report"



Select an item from the list to display that item's metadata form in the right-hand pane, click the current Item Type (typically either Journal Article or Document) to display the list of selections, select the "Report" type

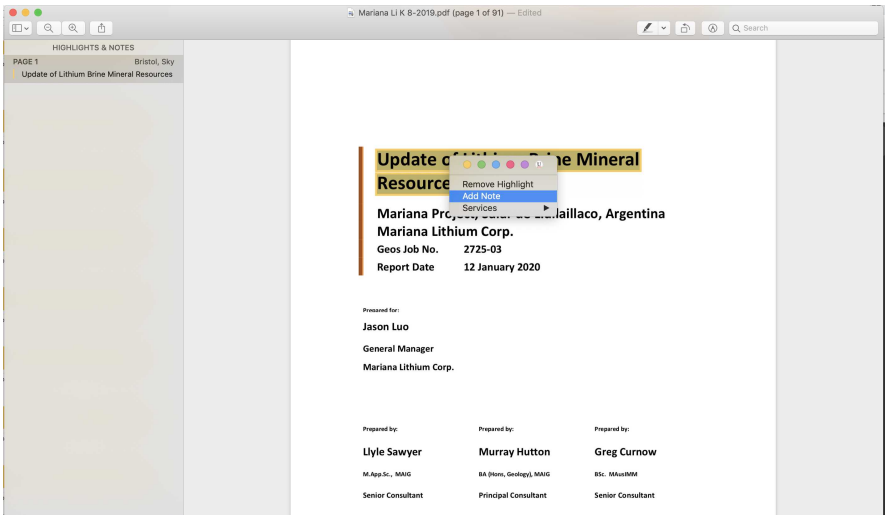
Annotate PDF Documents

4 We can now start annotating the PDFs documents stored with Zotero. Part of this will be developing a few fields of basic citation metadata. If the import process was able to successfully find things like title and authors, it is not necessary to annotate these elements separately. We already have them and can move that information into a shared library if we continue using Zotero in that capacity or export the information to include in another system. The following steps outline the approach to annotation that we can refine over time.

4.1 Open a PDF for annotation. Zotero should be set up to use your system default PDF reader. This will vary between operating systems and circumstances. You can change this setting in Zotero preferences if desired.

You can usually either double-click on the item in the Library or open the "twisty" next to the item to view its file and double click that to open.

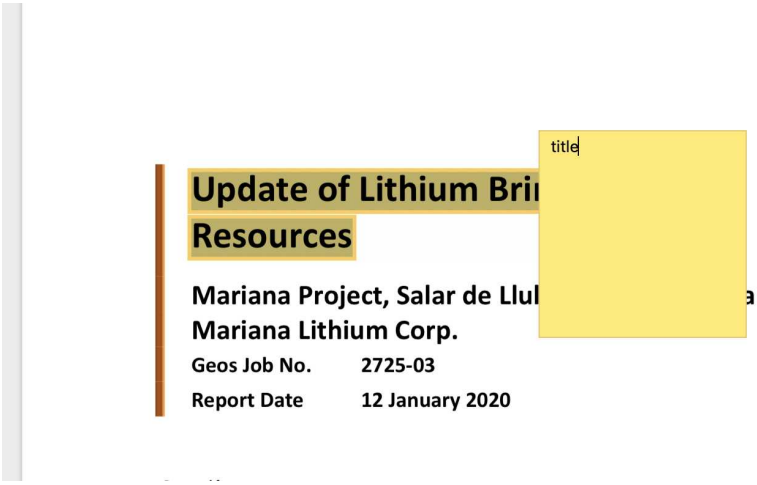
4.2 Highlight and annotate title (if necessary)



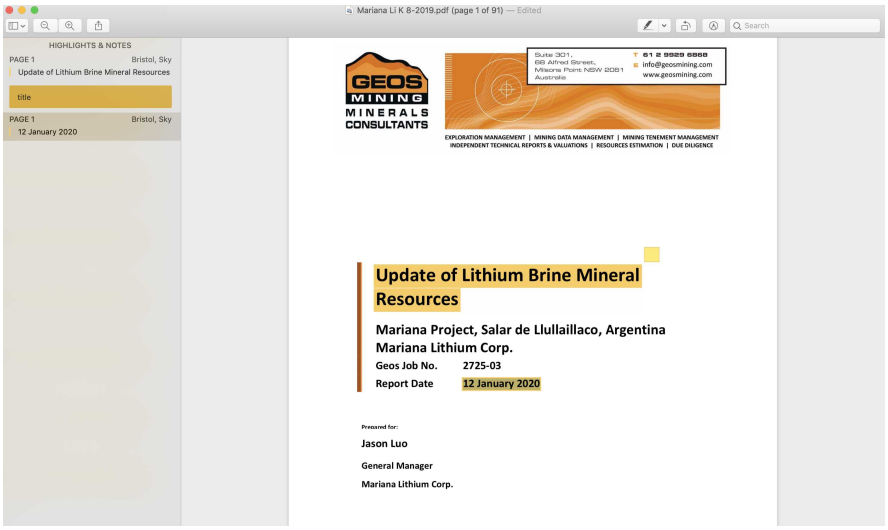
Select and highlight text for the title of the report that would be appropriate to list in a bibliographic reference. This will vary based on the report and can be found on a formatted title page or elsewhere in the document.

PDF reader software will vary on how to highlight a section of text, but this generally involves selecting text with a mouse and clicking some type of button or highlight feature. You can then add a note to that highlighted section. This will also vary and may involve right clicking the highlighted text or some other action to add a note.

Use **title** as the simple annotation for this element.



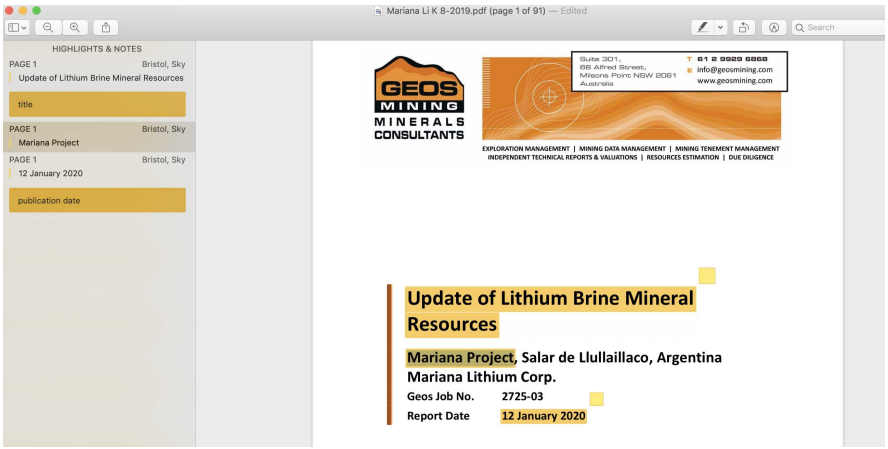
4.3 Highlight and annotate a date for the publication (if necessary)



The most appropriate date for the NI 43-101 Reports may be listed as "Effective Date" or something else like "Report Date" in this example. We need to identify a date to use as the publication date for the citation metadata.

Use **publication date** as the simple annotation for this element

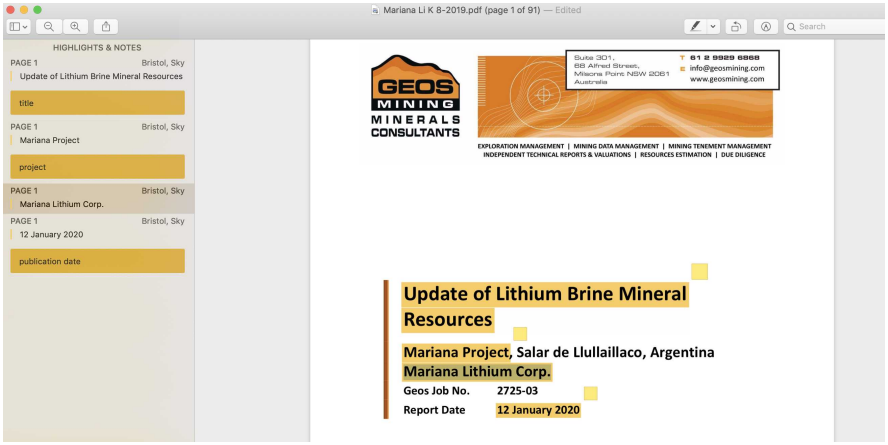
4.4 Highlight and annotate the project name



The project name may be found in different parts of the documents and may or may not include the word, project, in the text. We are looking for some type of simple shorthand reference to the particular mineral development project that can be used to identify sometimes multiple reports associated with the same project or can be found in other references. In some cases this may be the name of a mine or mining property that could be "well known" in other reference sources.

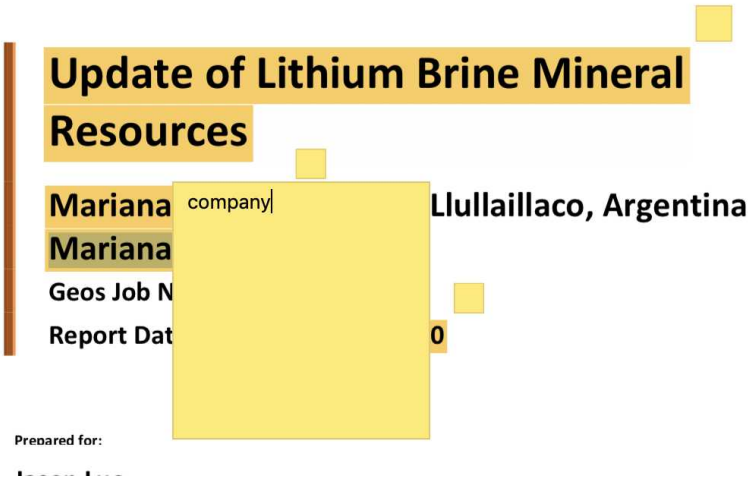
Use **project** as the simple annotation for this element

4.5 Highlight and annotate the company (or institution) that is operating the mineral development project



The name of the company operating the mine or exploration project may be listed in various places in the document.

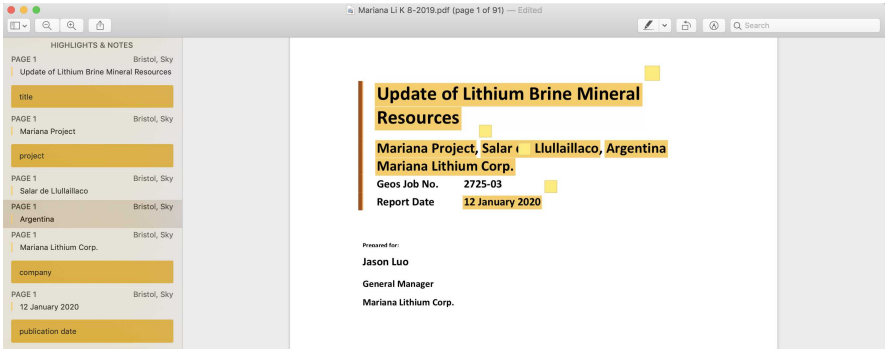
Use **company** as the simple annotation for this element



We may need to get more sophisticated at where and how we identify this and other company/organization names within the document. A standard named entity recognition method is identifying people and organizations within a corpus of text, and we may want to train these to identify the specific types of organizations we care about.

For now, the "company" name will be used to populate a metadata element called institution for building citation strings.

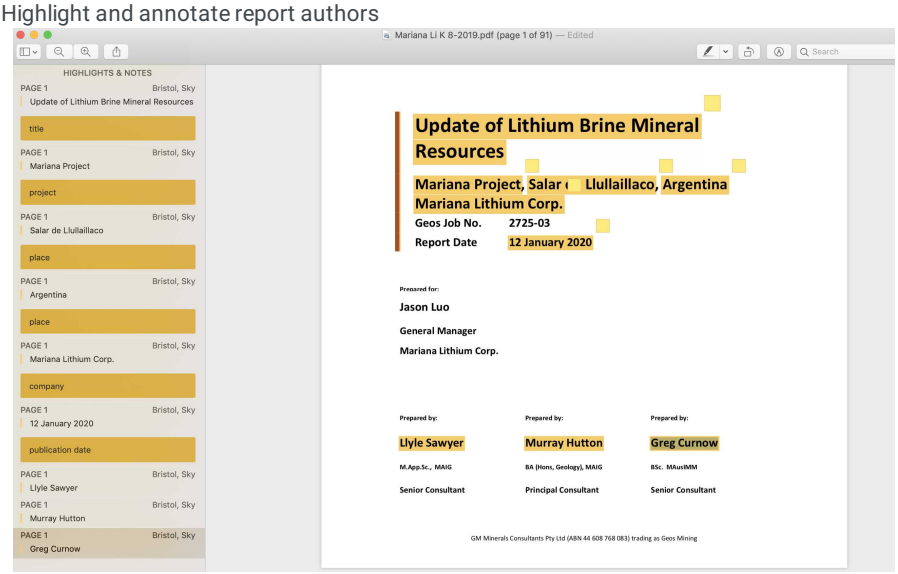
4.6 Highlight and annotate significant place names



Basic geographic context for the projects can be found in various places in these reports. What we are looking for are a few simple things like country names, provinces/states, counties, and other administrative units that can be used to generally locate the area where a mining property exists. These have been used to create a folder structure in the past, but we can pull them out as place name keywords as an alternative.

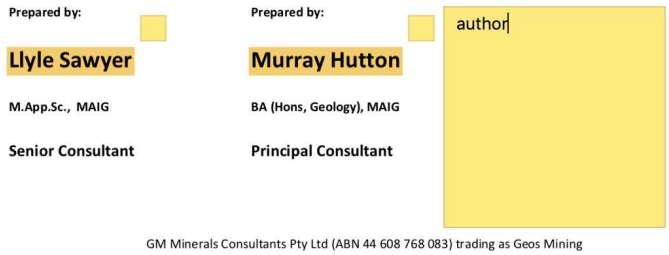
Use **place** as the simple annotation for this element

4.7 Highlight and annotate report authors



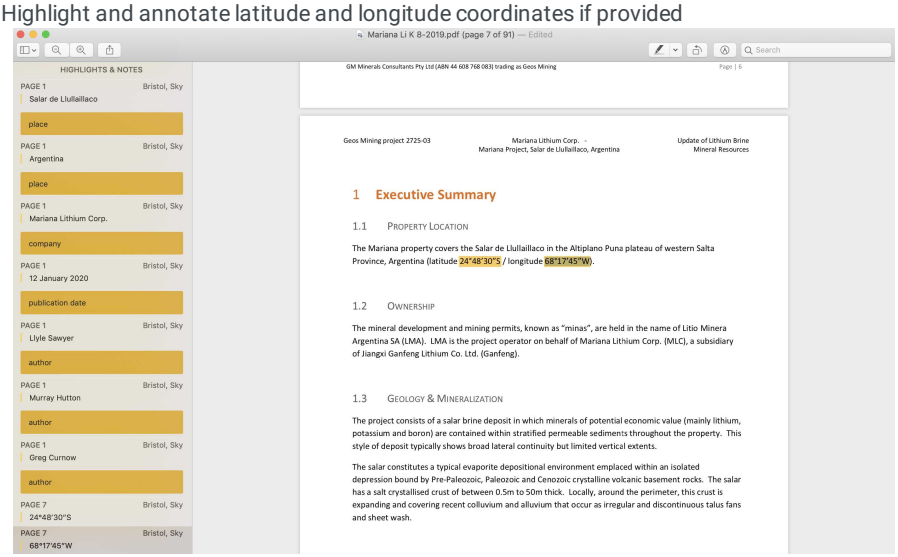
Author names can be listed in various ways in the NI 43-101 Reports. We are looking for the best reasonable text to use in identifying report authors for citation metadata.

Use author as the simple annotation for this element



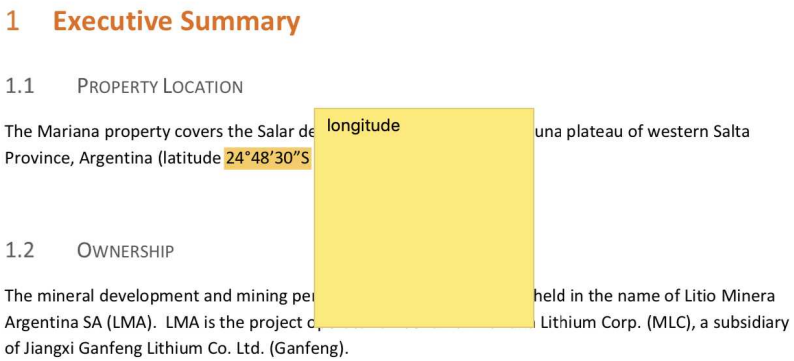
When extracting these annotations for use, multiple highlights using the same annotation will be pulled out in an ordered list based on where they appear in the document, giving us the ability to produce an ordered list of authors.

4.8 Highlight and annotate latitude and longitude coordinates if provided

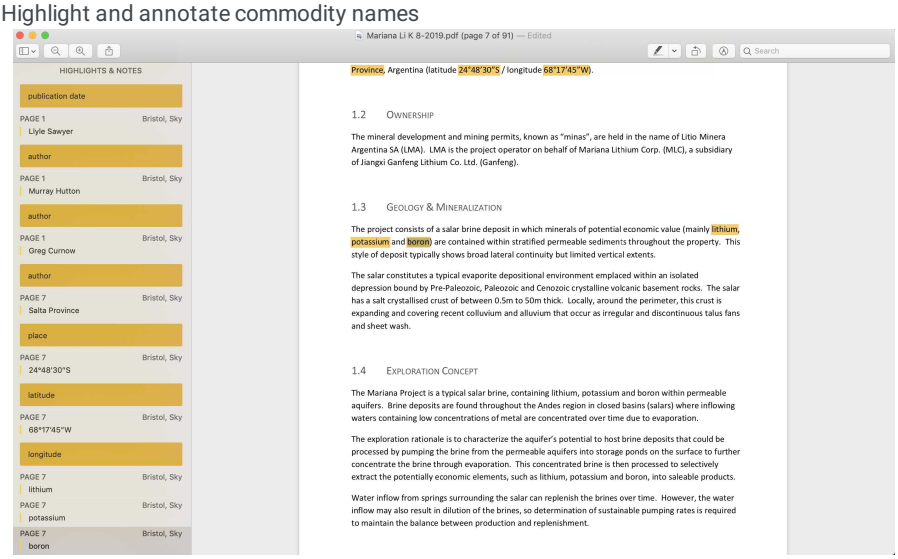


We ultimately want to identify point coordinates or more sophisticated geospatial footprints for as many of the mining properties as we can. These may be listed in the document in various places, and if we can identify patterns in how these are documented, we can potentially develop extraction algorithms. In the near term, highlight whatever form the coordinates are in, and we can extract and process those further.

Use **latitude** and **longitude** as the annotations for these elements

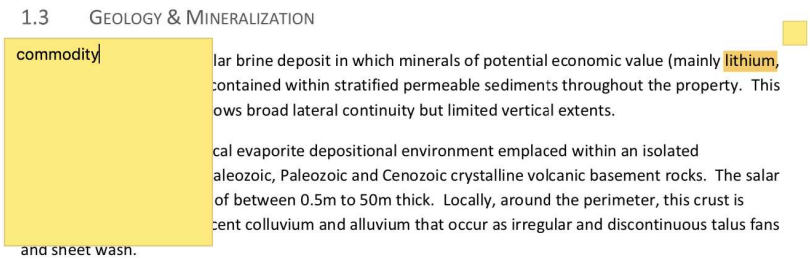


4.9 Highlight and annotate commodity names



Commodity names may be listed and summarized in different places in the document but are generally mentioned by name in the Geology & Mineralization section. In the near term, we need a simple identification of mineral commodities by name to bring into tags for helping to navigate to reports applicable for a given assessment or analysis. Longer term, we will work to use named entity recognition algorithms to identify any and all minerals that may be of interest, using natural language processing to determine their relative "significance" in terms of what is being described in a given text.

Use **commodity** as the simple annotation for this element



Saving and Processing

5 Save the PDF file at the end of or during annotation. Depending on the platform and software used, this may be done automatically. The highlights and annotations will be saved with the Zotero copy of the file.

At this stage, the most important thing is simply to get the annotations started within the files. Don't worry that the metadata you've highlighted is not a part of the citation information in Zotero yet. We'll get to that point shortly through a software process that will need to extract the annotations from the texts and place it back into Zotero. Right now, this is working against a shared, group library that is on Zotero's cloud, and we have to work out the best way of moving that to an operational state.