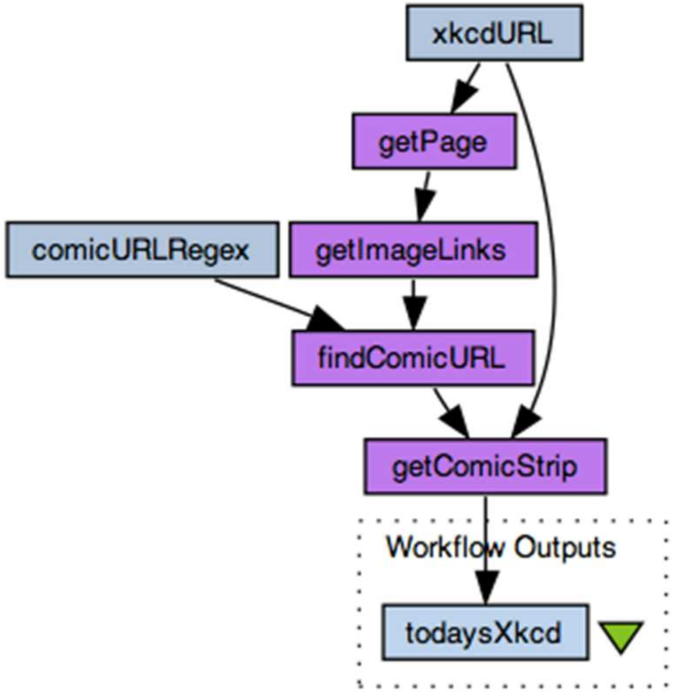
Trends in Use of Scientific Workflows: Insights from a Public Repository and Recommendations for Best Practices

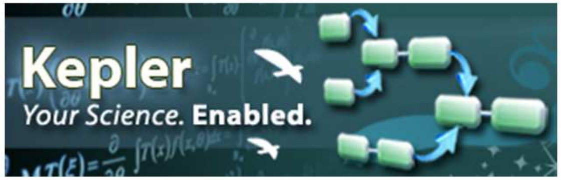
Richard Littauer, Karthik Ram, Bertram Ludäscher, William Michener, Rebecca Koskela

Scientific Workflows

- Tools that help scientists:
  - Automate repetitive or difficult work
  - Provide reproducibility to their experiments
  - Track provenance
  - Share their data with other scientists

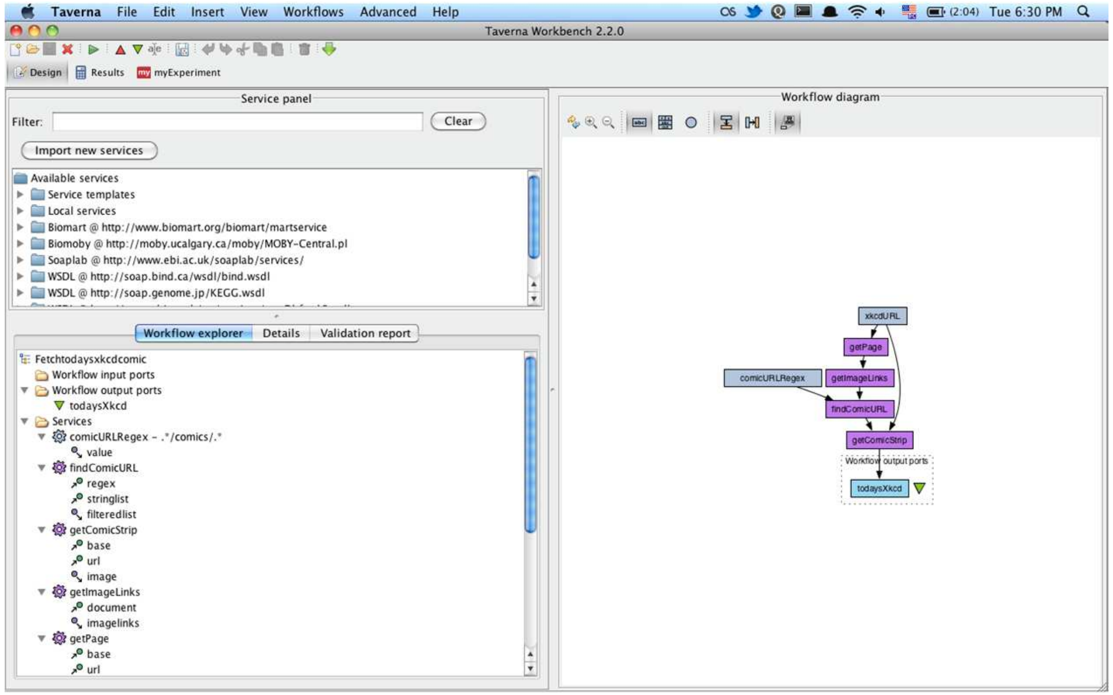


Workflow Workbenches

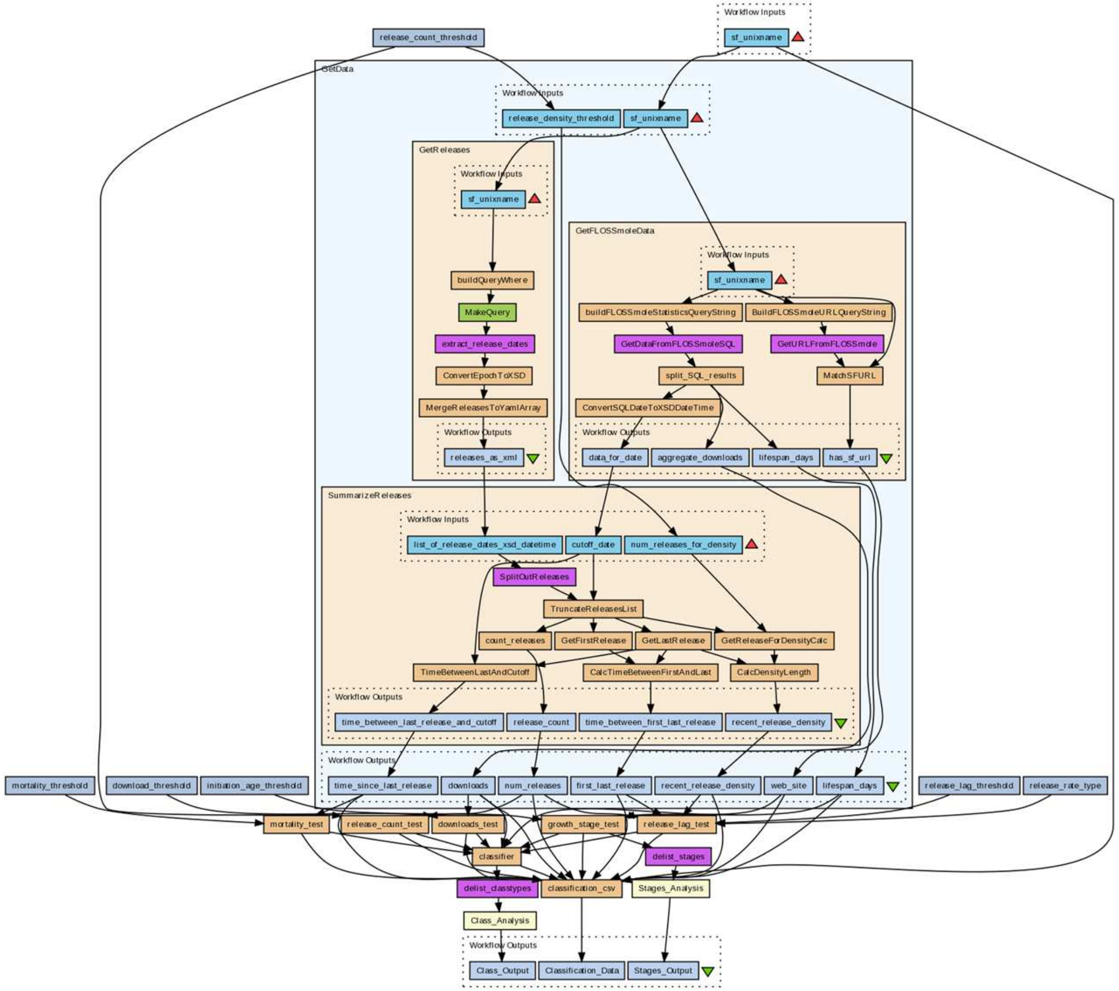


Workflow Workbenches

- These facilitate:
  - Creation
  - Mapping
  - Scheduling
  - Execution
  - Visualization
  - Re-Use

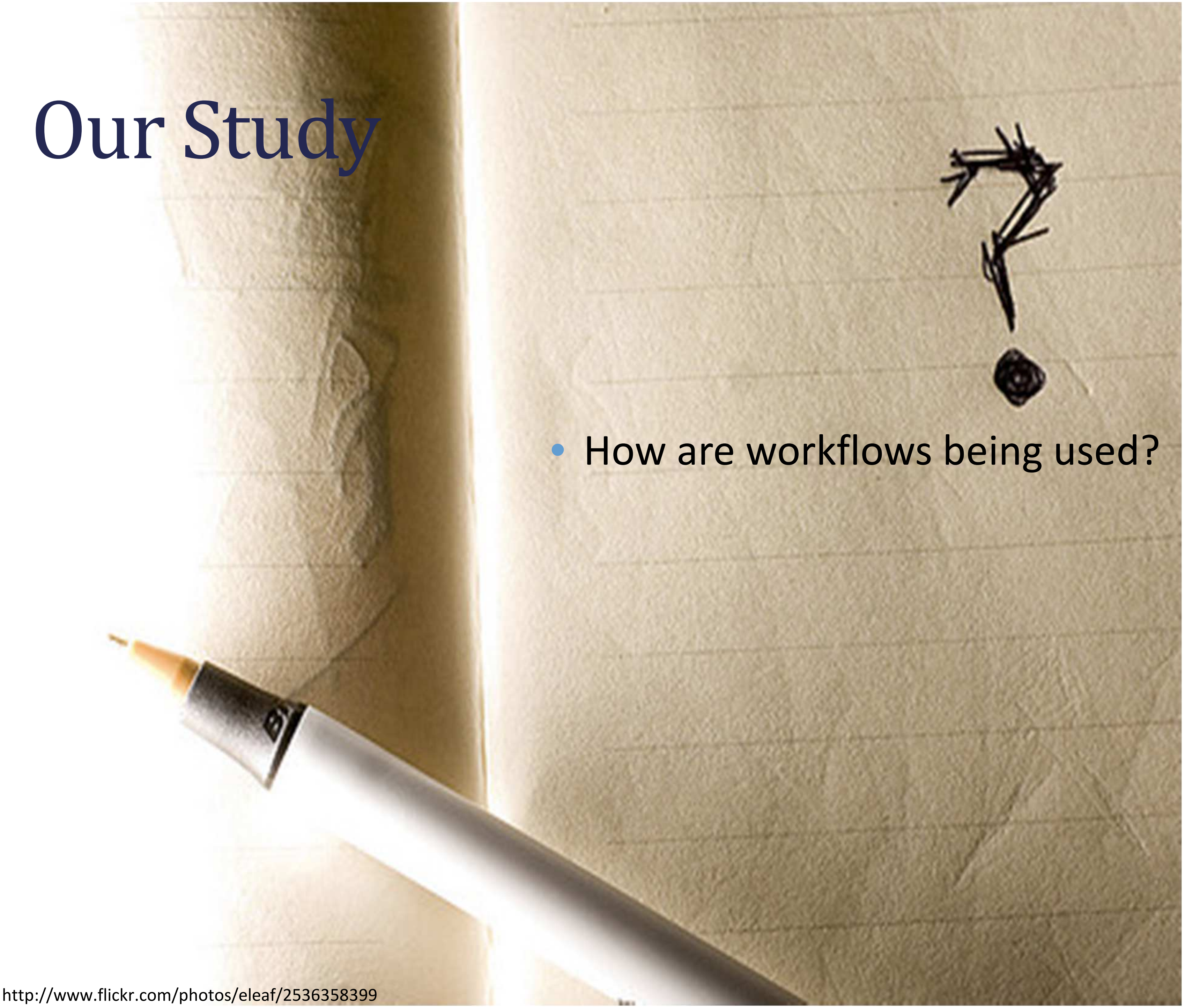


Example Workflow



http://www.myexperiment.org/workflows/140.html

DataONE



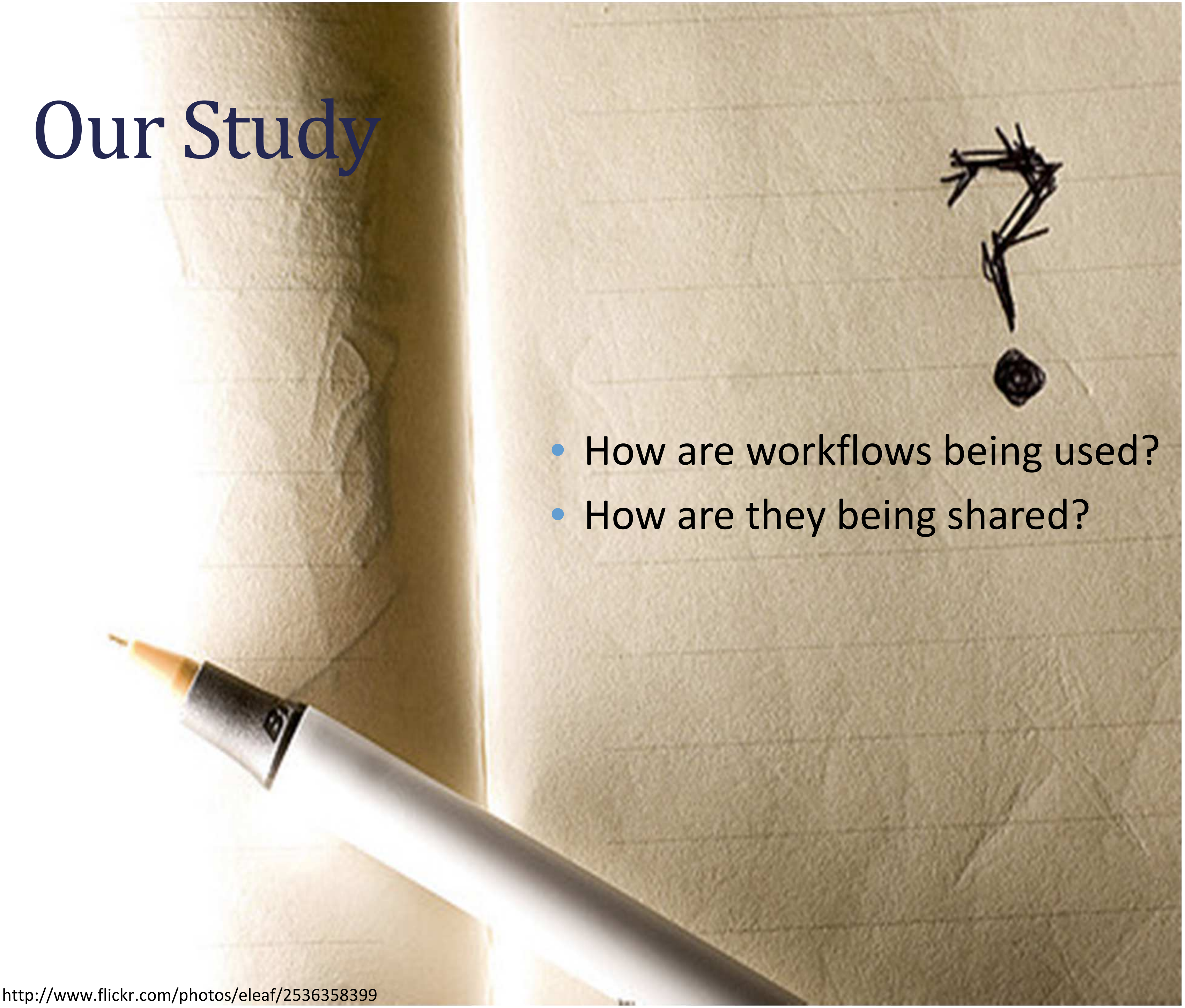
Our Study

- How are workflows being used?

http://www.flickr.com/photos/eleaf/2536358399

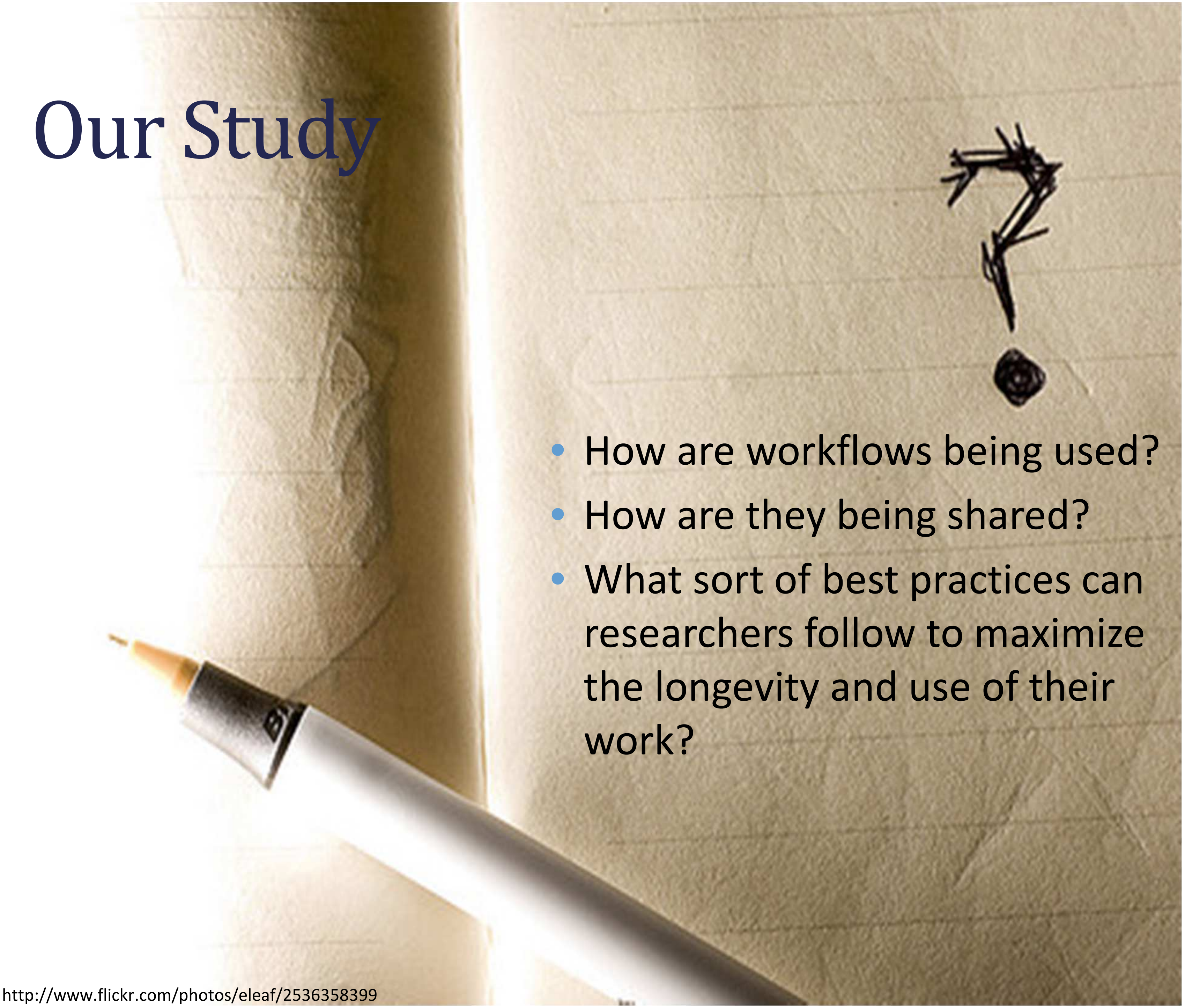
Our Study

- How are workflows being used?
- How are they being shared?



http://www.flickr.com/photos/eleaf/2536358399

Our Study



- How are workflows being used?
- How are they being shared?
- What sort of best practices can researchers follow to maximize the longevity and use of their work?

http://www.flickr.com/photos/eleaf/2536358399

Our Study

my experiment

- http://www.myexperiment.org
  - Est. 2007
  - 5000+ users
  - 2000+ workflows (mostly Taverna 1, 2, and RapidMiner)

Our Study

- http://www.myexperiment.org
  - Est. 2007
  - 5000+ users
  - 2000+ workflows (mostly Taverna 1, 2, and RapidMiner)

  - Minable RDF storage for workflows, groups, packs, users, files.
  - Minable data gathered through the SCUFLE XML language for the Taverna workflows
  - Taverna 1 - 479 workflows; Taverna 2 - 684 workflows.

Our Study

myexperiment

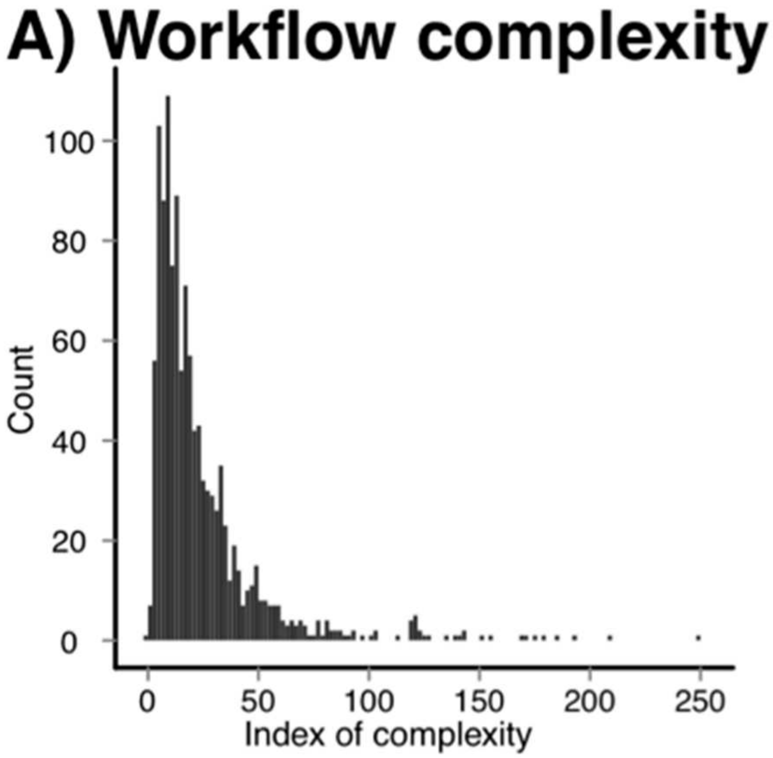
- We harvested information using a combination of SPARQL and Python (https://github.com/RichardLitt/Understanding-Workflows)

Our Study

my experiment

- We harvested information using a combination of SPARQL and Python (https://github.com/RichardLitt/Understanding-Workflows)
- Gathered user, workflow, files, packs, groups view and download statistics, metadata, descriptions, tags, and so on (http://thedatahub.org/dataset/myexperiment-screenscrape)

Findings

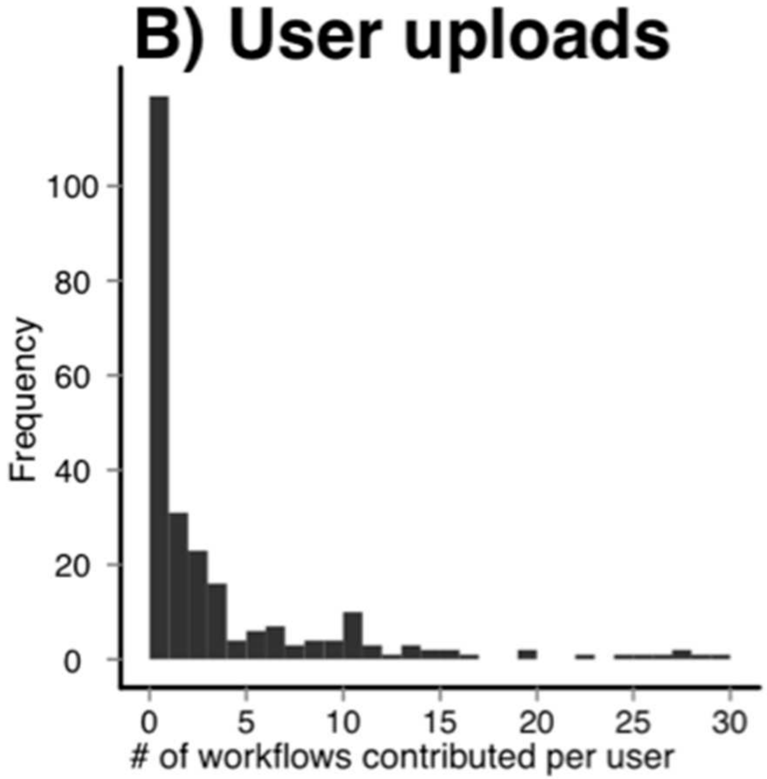


- A large percentage of workflows consist of *few components.*

- The amount of components ranges from 1 to 250. The average workflow supports 24.3 tasks.

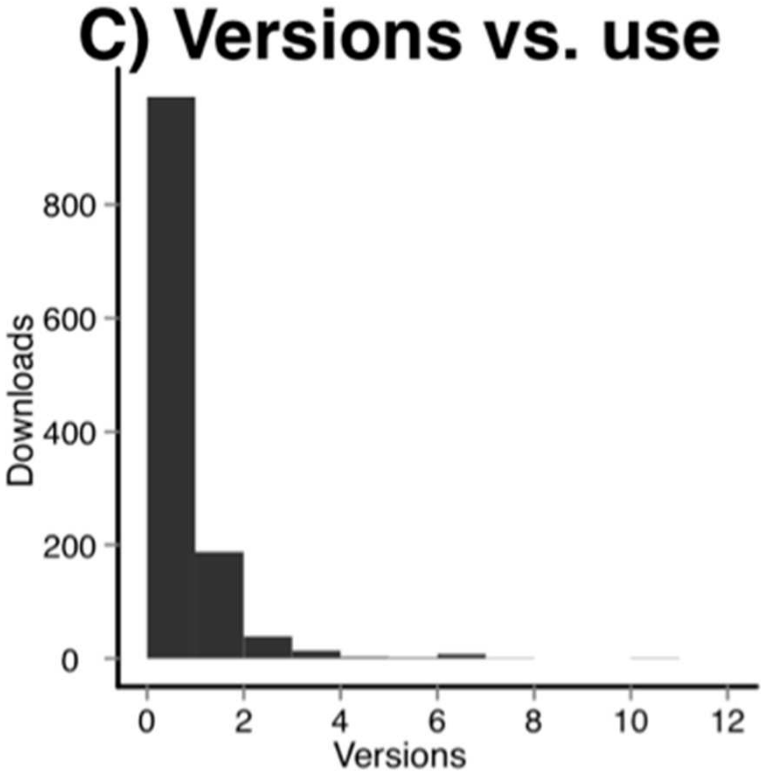
- Complex workflows that perform many tasks are downloaded more frequently than simpler workflows.

Findings



- Most workflow contributors submit a *single* workflow.
- Only 13 users have uploaded more than 30 workflows.
- Just over 5% of the users on myExperiment have uploaded workflows.

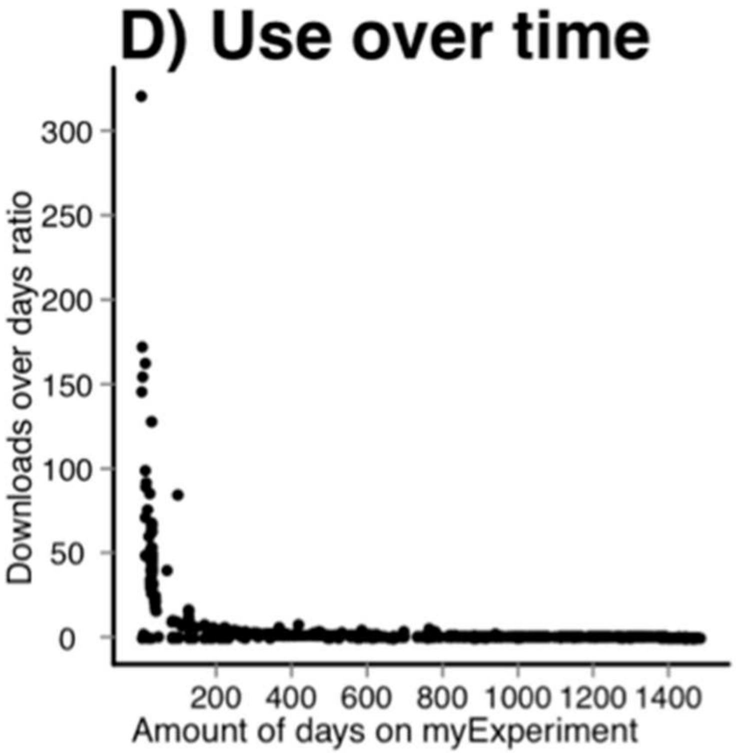
Findings



- Most workflows have only one version uploaded.

- When several versions do exist, the workflow is more frequently downloaded than “single-edition” workflows.

Findings



- Workflow use declined significantly a month after initial upload.

Findings

- A large percentage of workflow components – approx. 38% - are *shims.*
  - Components that are used to make output from one step conform to the format expected by a subsequent step.

Findings

- A large percentage of workflow components – approx. 38% - are *shims.*
  - Components that are used to make output from one step conform to the format expected by a subsequent step.
  - This is a problem for developers.

Findings

- A large percentage of workflow components – approx. 38% - are *shims.*
  - Components that are used to make output from one step conform to the format expected by a subsequent step.
  - This is a problem for developers.
  - 8% more than previous studies [1]

Findings

- 60% of workflows have *embedded workflows* within them.

Findings

- 60% of workflows have *embedded workflows* within them.

- Use of workflows does **not** seem to be related to the volume of documentation associated with the workflow nor the number of tags…

Findings

- 60% of workflows have *embedded workflows* within them.
- Use of workflows does **not** seem to be related to the volume of documentation associated with the workflow nor the number of tags…
- … but **is** related to the degree of *community engagement* with the workflow as exhibited by number of citations, comments, ratings, and reviews.

Recommendations

Remember workflows are evolving entities.

They are updated in response to user feedback, engagement, and improvements in methodology.

Recommendations

Use relevant social annotation tools.

But they need to be constrained; for instance, through the use of a controlled tag vocabulary.

Recommendations

Talk about them.

Cite the workflow in publications.

Share with colleagues

Advertise the workflow.

Recommendations

Provide sufficient descriptions of your workflows.

Recommendations

Keep in mind that one size does not fit all.

Recommendations

Workflow re-use could benefit significantly from the assignment of stable identifiers, like Digital Object Identifiers (DOI).

Recommendations

Increased usage of workflows and workflow repositories will likely be related to the degree that education is provided to scientists through professional society meetings, online courses, and incorporation into academic training (e.g. undergraduate and graduate courses).

Impact on Science

Following these recommendations can help:

- Make science more efficient.

Impact on Science

Following these recommendations can help:

- Make science more efficient.
- Promote reproducible science.

Impact on Science

Following these recommendations can help:

- Make science more efficient.
- Promote reproducible science.
- Provide further venues for cataloging an individual's research contributions.

Impact on Science

Following these recommendations can help:

- Make science more efficient.
- Promote reproducible science.
- Provide further venues for cataloging an individual's research contributions.
- Speed up the peer review process.

Impact on Science

Following these recommendations can help:

- Make science more efficient.
- Promote reproducible science.
- Provide further venues for cataloging an individual's research contributions.
- Speed up the peer review process.
- Your impact; NSF considers workflows to be valuable contributions.

References

Kepler Project. http://www.kepler-project.org

Taverna. http://www.taverna.org.uk/

- [1] Cui Lin, Shiyong Lu, Xubo Fei, Darshan Pai, and Jing Hua. 2009. A Task Abstraction and Mapping Approach to the Shimming Problem in Scientific Workflows. In *Proceedings of the 2009 IEEE International Conference on Services Computing* (SCC '09). IEEE Computer Society, Washington, DC, USA, http://dx.doi.org/10.1109/SCC.2009.77

http://www.flickr.com/photos/wwworks/4759535950/

Links

- Mendeley Research Group: http://www.mendeley.com/groups/1189721/scientific-workflows-and-workflow-systems/
- Github https://github.com/RichardLitt/Understanding-Workflows
- Data http://thedatahub.org/dataset/myexperiment-screenscrape
- Notebook https://notebooks.dataone.org/workflows

http://www.flickr.com/photos/wwworks/4759535950/