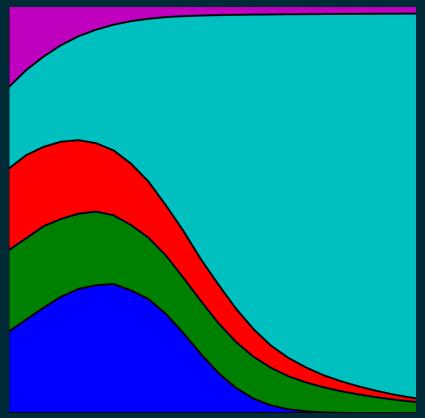
Arcas: Using Python to access open research literature

@NikoletaGlyn

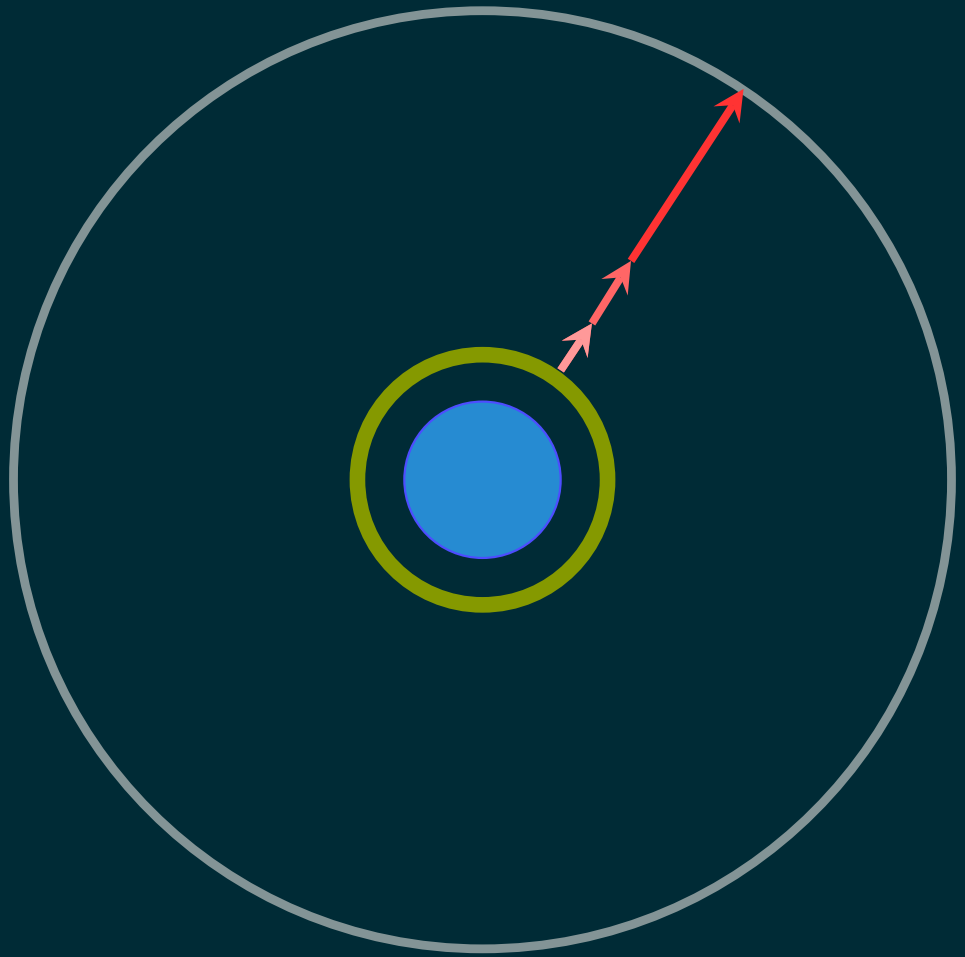


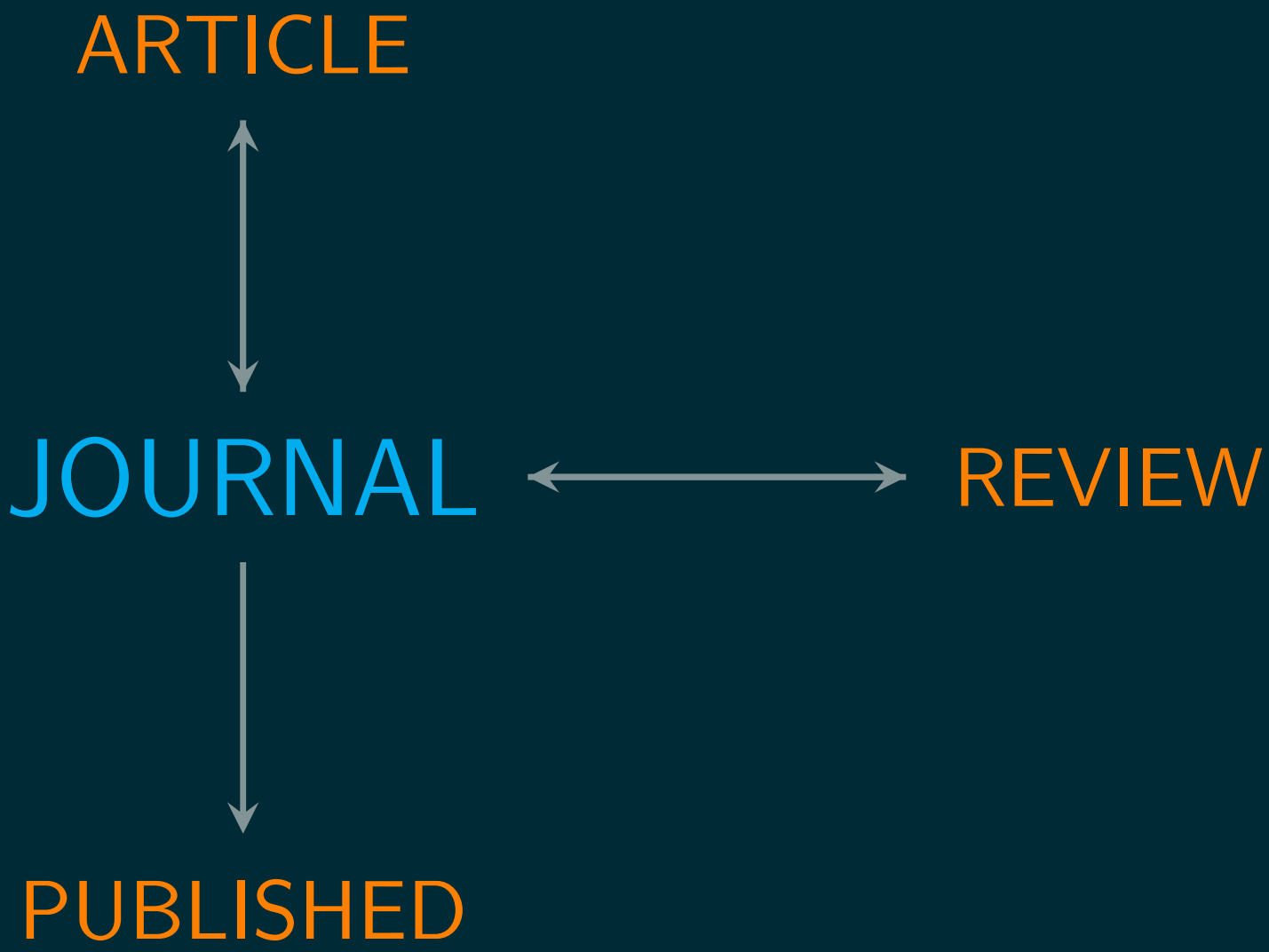
Software
Sustainability
Institute

The illustrated guide to a Ph.D.

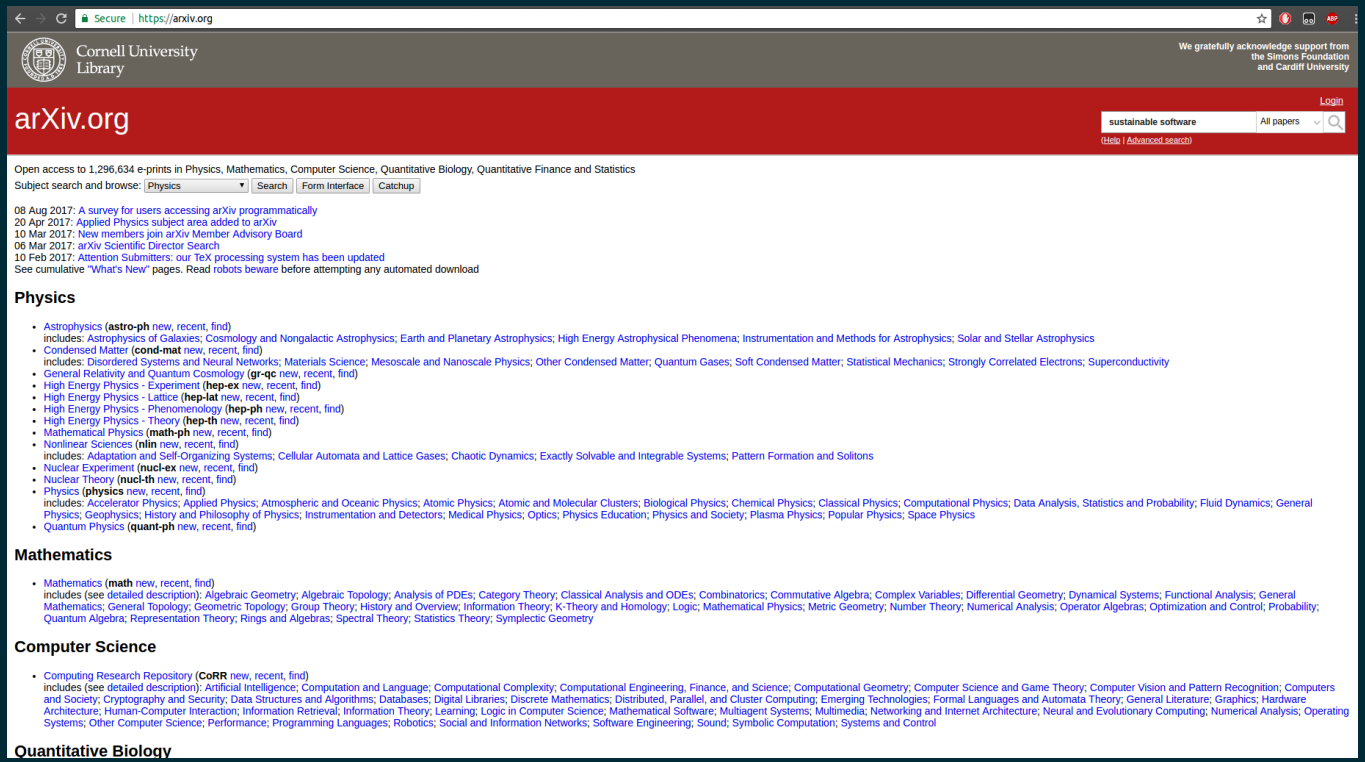
Matt Might

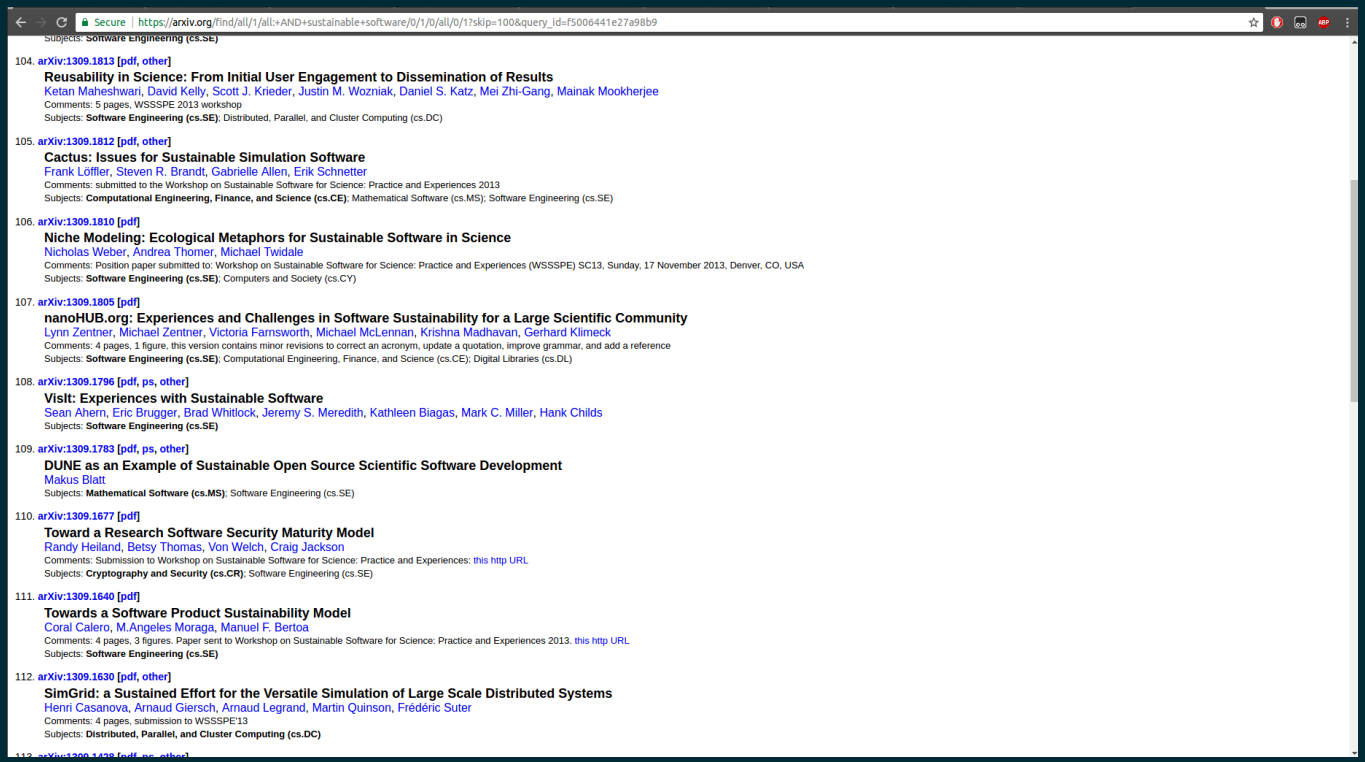
http://matt.might.net/articles/phd-school-in-pictures/




ARTICLE
JOURNAL
REVIEW
PUBLISHED

Sustainable Software


Secure | https://arxiv.org
Cornell University Library
We gratefully acknowledge support from the Simons Foundation and Cardiff University
arXiv.org
Login
sustainable software
All papers
(Help | Advanced search)
Open access to 1,296,634 e-prints in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance and Statistics
Subject search and browse: Physics
Search
Form Interface
Catchup
08 Aug 2017: A survey for users accessing arXiv programmatically
20 Apr 2017: Applied Physics subject area added to arXiv
10 Mar 2017: New members join arXiv Member Advisory Board
06 Mar 2017: arXiv Scientific Director Search
10 Feb 2017: Attention Submitters: our TeX processing system has been updated
See cumulative "What's New" pages. Read robots beware before attempting any automated download
Physics
Astrophysics (astro-ph new, recent, find)
includes: Astrophysics of Galaxies; Cosmology and Nongalactic Astrophysics; Earth and Planetary Astrophysics; High Energy Astrophysical Phenomena; Instrumentation and Methods for Astrophysics; Solar and Stellar Astrophysics
Condensed Matter (cond-mat new, recent, find)
includes: Disordered Systems and Neural Networks; Materials Science; Mesoscale and Nanoscale Physics; Other Condensed Matter; Quantum Gases; Soft Condensed Matter; Statistical Mechanics; Strongly Correlated Electrons; Superconductivity
General Relativity and Quantum Cosmology (gr-qc new, recent, find)
High Energy Physics - Experiment (hep-ex new, recent, find)
High Energy Physics - Lattice (hep-lat new, recent, find)
High Energy Physics - Phenomenology (hep-ph new, recent, find)
High Energy Physics - Theory (hep-th new, recent, find)
Mathematical Physics (math-ph new, recent, find)
Nonlinear Sciences (nlin new, recent, find)
includes: Adaptation and Self-Organizing Systems; Cellular Automata and Lattice Gases; Chaotic Dynamics; Exactly Solvable and Integrable Systems; Pattern Formation and Solitons
Nuclear Experiment (nucl-ex new, recent, find)
Nuclear Theory (nucl-th new, recent, find)
Physics (physics new, recent, find)
includes: Accelerator Physics; Applied Physics; Atmospheric and Oceanic Physics; Atomic Physics; Atomic and Molecular Clusters; Biological Physics; Chemical Physics; Classical Physics; Computational Physics; Data Analysis, Statistics and Probability; Fluid Dynamics; General Physics; Geophysics; History and Philosophy of Physics; Instrumentation and Detectors; Medical Physics; Optics; Physics Education; Physics and Society; Plasma Physics; Popular Physics; Space Physics
Quantum Physics (quant-ph new, recent, find)
Mathematics
Mathematics (math new, recent, find)
includes (see detailed description): Algebraic Geometry; Algebraic Topology; Analysis of PDEs; Category Theory; Classical Analysis and ODEs; Combinatorics; Commutative Algebra; Complex Variables; Differential Geometry; Dynamical Systems; Functional Analysis; General Mathematics; General Topology; Geometric Topology; Group Theory; History and Overview; Information Theory; K-Theory and Homology; Logic; Mathematical Physics; Metric Geometry; Number Theory; Numerical Analysis; Operator Algebras; Optimization and Control; Probability; Quantum Algebra; Representation Theory; Rings and Algebras; Spectral Theory; Statistics Theory; Symplectic Geometry
Computer Science
Computing Research Repository (CoRR new, recent, find)
includes (see detailed description): Artificial Intelligence; Computation and Language; Computational Complexity; Computational Engineering, Finance, and Science; Computational Geometry; Computer Science and Game Theory; Computer Vision and Pattern Recognition; Computers and Society; Cryptography and Security; Data Structures and Algorithms; Databases; Digital Libraries; Discrete Mathematics; Distributed, Parallel, and Cluster Computing; Emerging Technologies; Formal Languages and Automata Theory; General Literature; Graphics; Hardware Architecture; Human-Computer Interaction; Information Retrieval; Information Theory; Learning; Logic in Computer Science; Mathematical Software; Multiagent Systems; Multimedia; Networking and Internet Architecture; Neural and Evolutionary Computing; Numerical Analysis; Operating Systems; Other Computer Science; Performance; Programming Languages; Robotics; Social and Information Networks; Software Engineering; Sound; Symbolic Computation; Systems and Control
Quantitative Biology


Secure | https://arxiv.org/find/all/1/all:+AND+sustainable+software/0/1/0/all/0/1?skip=100&query_id=f5006441e27a98b9
Subjects: Software Engineering (cs.SE)
104. arXiv:1309.1813 [pdf, other]
Reusability in Science: From Initial User Engagement to Dissemination of Results
Ketan Maheshwari, David Kelly, Scott J. Krieder, Justin M. Wozniak, Daniel S. Katz, Mei Zhi-Gang, Mainak Mookherjee
Comments: 5 pages, WSSSPE 2013 workshop
Subjects: Software Engineering (cs.SE); Distributed, Parallel, and Cluster Computing (cs.DC)
105. arXiv:1309.1812 [pdf, other]
Cactus: Issues for Sustainable Simulation Software
Frank Löffler, Steven R. Brandt, Gabrielle Allen, Erik Schnetter
Comments: submitted to the Workshop on Sustainable Software for Science: Practice and Experiences 2013
Subjects: Computational Engineering, Finance, and Science (cs.CE); Mathematical Software (cs.MS); Software Engineering (cs.SE)
106. arXiv:1309.1810 [pdf]
Niche Modeling: Ecological Metaphors for Sustainable Software in Science
Nicholas Weber, Andrea Thomer, Michael Twidale
Comments: Position paper submitted to: Workshop on Sustainable Software for Science: Practice and Experiences (WSSSPE) SC13, Sunday, 17 November 2013, Denver, CO, USA
Subjects: Software Engineering (cs.SE); Computers and Society (cs.CY)
107. arXiv:1309.1805 [pdf]
nanoHUB.org: Experiences and Challenges in Software Sustainability for a Large Scientific Community
Lynn Zentner, Michael Zentner, Victoria Farnsworth, Michael McLennan, Krishna Madhavan, Gerhard Klimeck
Comments: 4 pages, 1 figure, this version contains minor revisions to correct an acronym, update a quotation, improve grammar, and add a reference
Subjects: Software Engineering (cs.SE); Computational Engineering, Finance, and Science (cs.CE); Digital Libraries (cs.DL)
108. arXiv:1309.1796 [pdf, ps, other]
VisIt: Experiences with Sustainable Software
Sean Ahern, Eric Brugger, Brad Whitlock, Jeremy S. Meredith, Kathleen Biagas, Mark C. Miller, Hank Childs
Subjects: Software Engineering (cs.SE)
109. arXiv:1309.1783 [pdf, ps, other]
DUNE as an Example of Sustainable Open Source Scientific Software Development
Makus Blatt
Subjects: Mathematical Software (cs.MS); Software Engineering (cs.SE)
110. arXiv:1309.1677 [pdf]
Toward a Research Software Security Maturity Model
Randy Heiland, Betsy Thomas, Von Welch, Craig Jackson
Comments: Submission to Workshop on Sustainable Software for Science: Practice and Experiences: this http URL
Subjects: Cryptography and Security (cs.CR); Software Engineering (cs.SE)
111. arXiv:1309.1640 [pdf]
Towards a Software Product Sustainability Model
Coral Calero, M.Angeles Moraga, Manuel F. Bertoa
Comments: 4 pages, 3 figures. Paper sent to Workshop on Sustainable Software for Science: Practice and Experiences 2013. this http URL
Subjects: Software Engineering (cs.SE)
112. arXiv:1309.1630 [pdf, other]
SimGrid: a Sustained Effort for the Versatile Simulation of Large Scale Distributed Systems
Henri Casanova, Arnaud Giersch, Arnaud Legrand, Martin Quinson, Frédéric Suter
Comments: 4 pages, submission to WSSSPE'13
Subjects: Distributed, Parallel, and Cluster Computing (cs.DC)

API

QUERY

http://export.arxiv.org/api/query?search_query=
Sustainable%20Software

export.arxiv.org/api/query?search_query=Visit%20Sustainable%20Software

```
<?xml version="1.0" encoding="UTF-8"?>
<feed xmlns="http://www.w3.org/2005/Atom">
  <link href="http://arxiv.org/api/query?search_query%3DVisit%20Sustainable%20Software%26id_list%3D%26start%3D0%26max_results%3D10" rel="self" type="application/atom+xml"/>
  <title type="html">ArXiv Query: search_query=Visit Sustainable Software&amp;amp;id_list=&amp;amp;start=0&amp;amp;max_results=10</title>
  <id>http://arxiv.org/api/0yoLYUVopsAfNXq+IKF9iORM6UU</id>
  <updated>2017-08-25T00:00:00-04:00</updated>
  <opensearch:totalResults xmlns:opensearch="http://a9.com/-/spec/opensearch/1.1/">19522</opensearch:totalResults>
  <opensearch:startIndex xmlns:opensearch="http://a9.com/-/spec/opensearch/1.1/">0</opensearch:startIndex>
  <opensearch:itemsPerPage xmlns:opensearch="http://a9.com/-/spec/opensearch/1.1/">10</opensearch:itemsPerPage>
  <entry>
    <id>http://arxiv.org/abs/1309.1796v1</id>
    <updated>2013-09-07T00:16:52Z</updated>
    <published>2013-09-07T00:16:52Z</published>
    <title>VisIt: Experiences with Sustainable Software</title>
    <summary>  The success of the VisIt visualization system has been wholly dependent upon
the culture and practices of software development that have fostered its
welcome by users and embrace by developers and researchers. In the following
paper, we, the founding developers and designers of VisIt, summarize some of
the major efforts, both successful and unsuccessful, that we have undertaken in
the last thirteen years to foster community, encourage research, create a
sustainable open-source development model, measure impact, and support
production software. We also provide commentary about the career paths that our
development work has engendered.
</summary>
    <author>
      <name>Sean Ahern</name>
    </author>
    <author>
      <name>Eric Brugger</name>
    </author>
    <author>
      <name>Brad Whitlock</name>
    </author>
    <author>
      <name>Jeremy S. Meredith</name>
    </author>
    <author>
      <name>Kathleen Biagas</name>
    </author>
    <author>
      <name>Mark C. Miller</name>
    </author>
    <author>
      <name>Hank Childs</name>
    </author>
    <link href="http://arxiv.org/abs/1309.1796v1" rel="alternate" type="text/html"/>
    <link title="pdf" href="http://arxiv.org/pdf/1309.1796v1" rel="related" type="application/pdf"/>
    <arxiv:primary_category xmlns:arxiv="http://arxiv.org/schemas/atom" term="cs.SE" scheme="http://arxiv.org/schemas/atom"/>
    <category term="cs.SE" scheme="http://arxiv.org/schemas/atom"/>
  </entry>
  <entry>
    <id>http://arxiv.org/abs/1608.04529v2</id>
    <updated>2016-08-17T10:29:58Z</updated>
    <published>2016-08-16T09:41:23Z</published>
    <title>A Proposal for the Measurement and Documentation of Research Software
  Sustainability in Interactive Metadata Repositories</title>
    <summary>  This paper proposes an interactive repository type for research software
metadata which measures and documents software sustainability by accumulating
metadata, and computing sustainability metrics over them. Such a repository
would help to overcome technical barriers to software sustainability by
furthering the discovery and identification of sustainable software, thereby
also facilitating documentation of research software within the framework of
software management plans.
</summary>
    <author>
```

QUERY

http://export.arxiv.org/api/query?search_query=
Sustainable%20Software

QUERY

http://export.arxiv.org/api/query?search_query=
Sustainable%20Software

http://api.plos.org/search?q=title:
Sustainable%20Software&rows=100

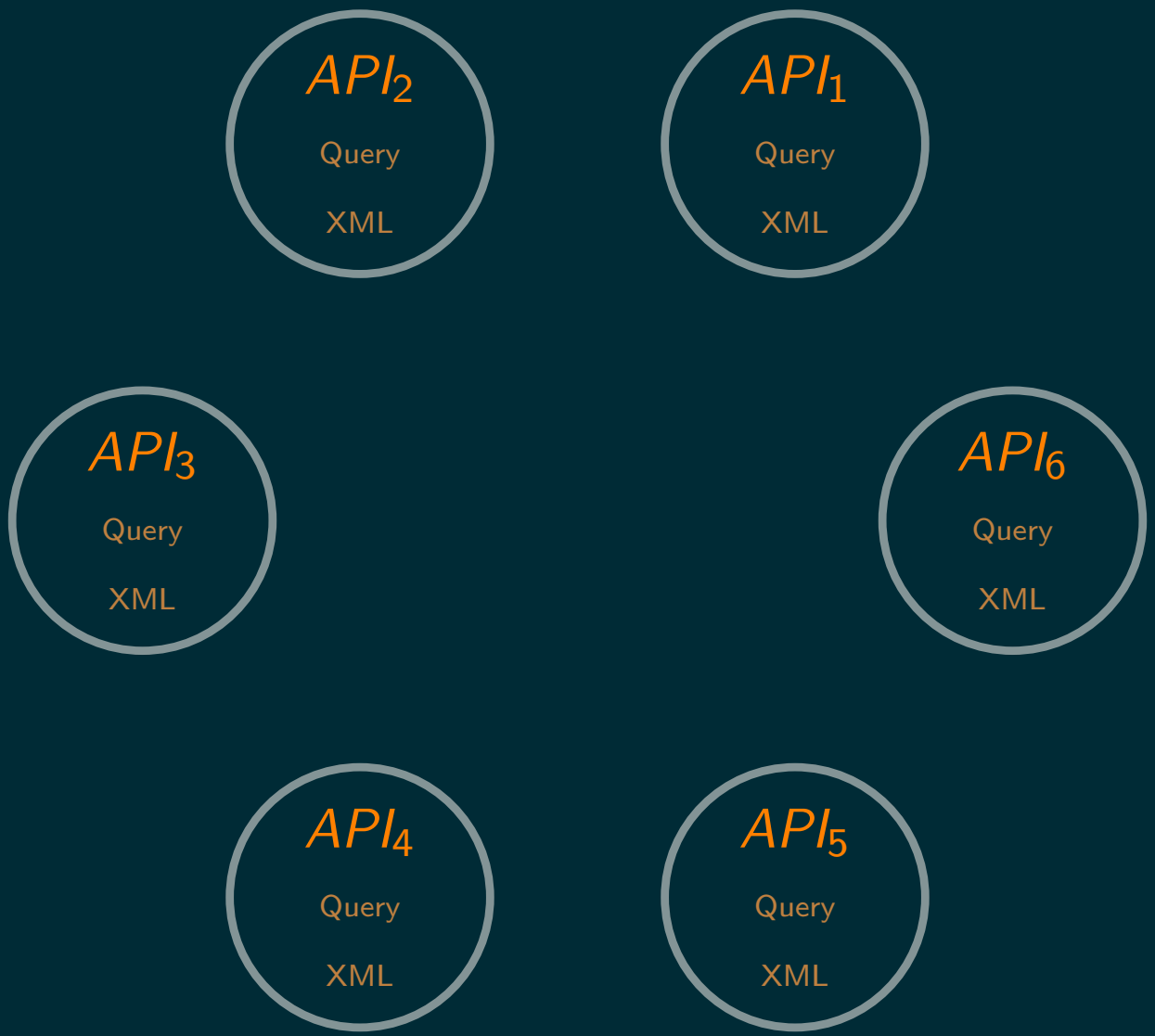
QUERY

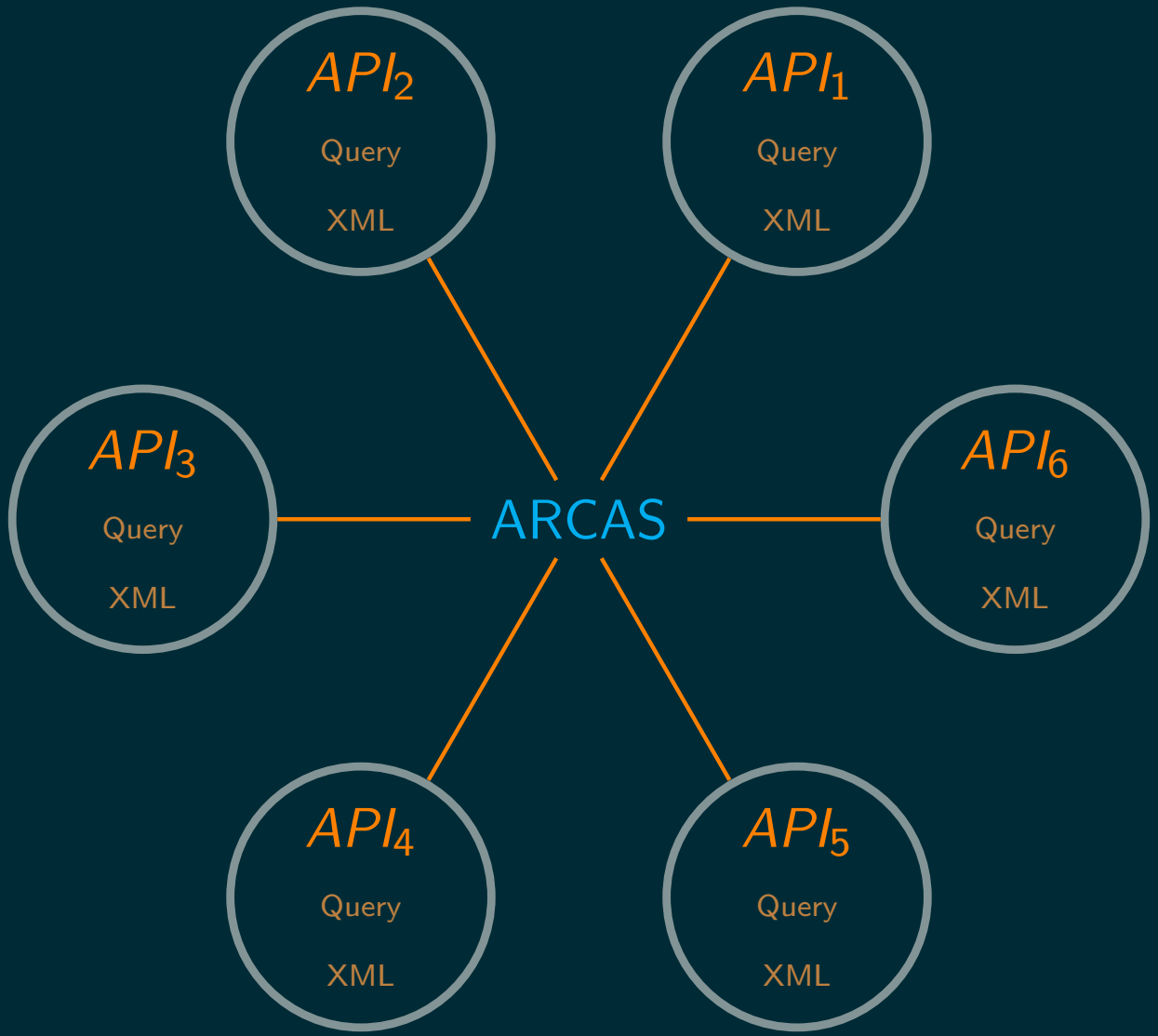
http://export.arxiv.org/api/query?search_query=
Sustainable%20Software

http://api.plos.org/search?q=title:
Sustainable%20Software&rows=100

http:
//www.nature.com/opensearch/request?queryType=cql&query=
dc.title%20adj%20SustainableSoftware&maximumRecords=100

...


API_2
Query
XML
API_1
Query
XML
API_3
Query
XML
API_6
Query
XML
API_4
Query
XML
API_5
Query
XML

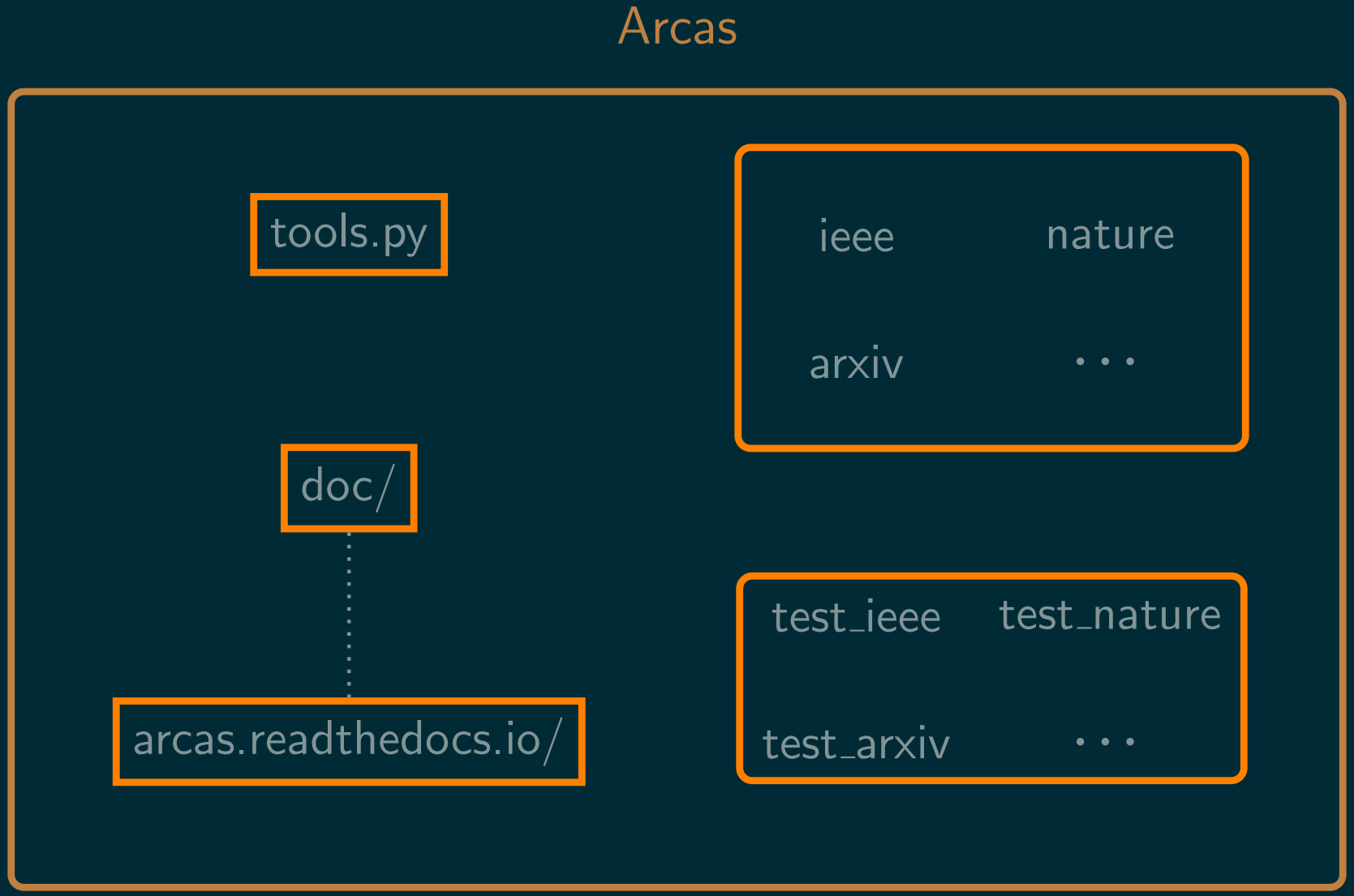

API_2
Query
XML
API_1
Query
XML
API_3
Query
XML
ARCAS
API_6
Query
XML
API_4
Query
XML
API_5
Query
XML

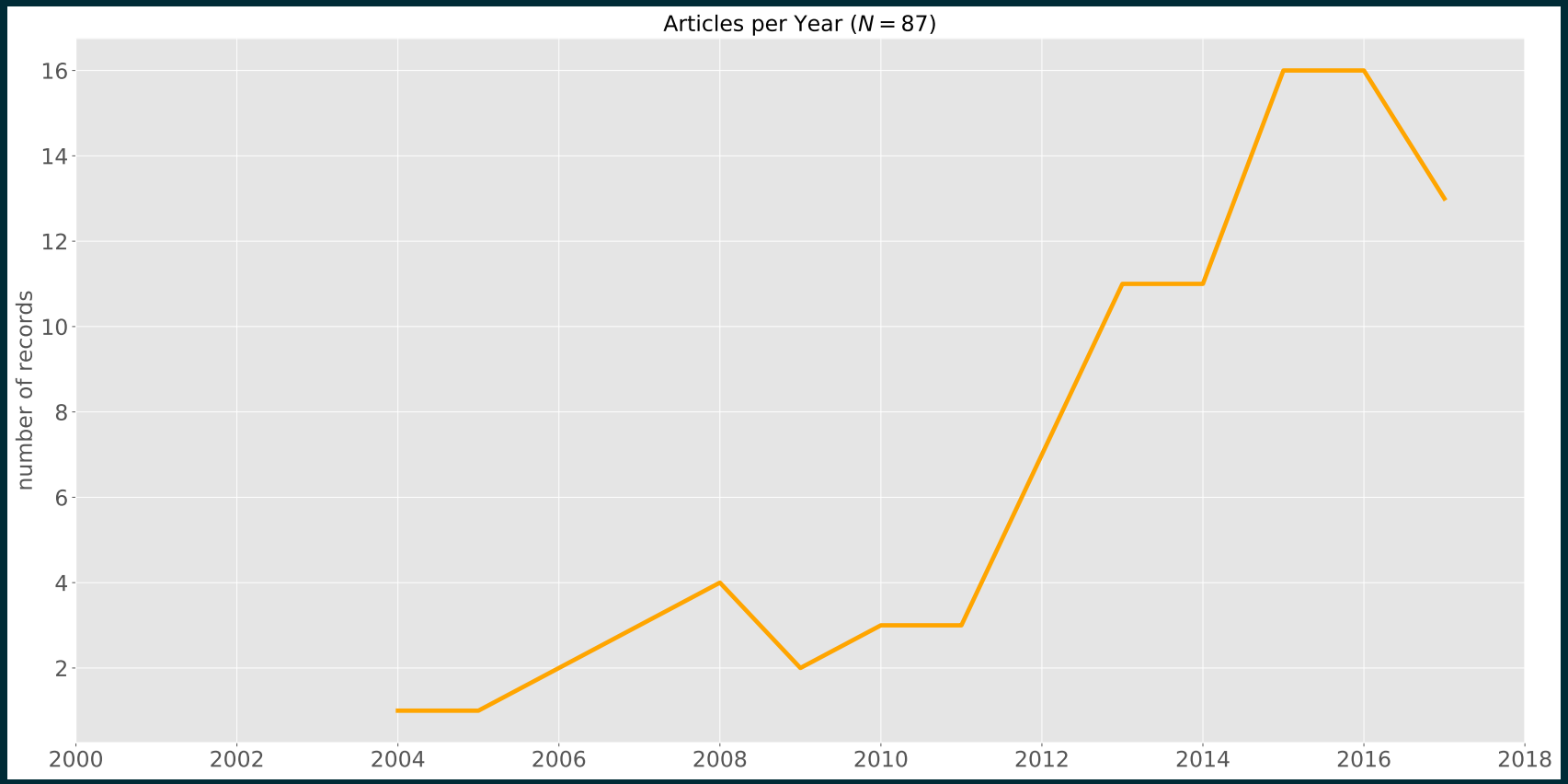
```
$ pip install arcas
```

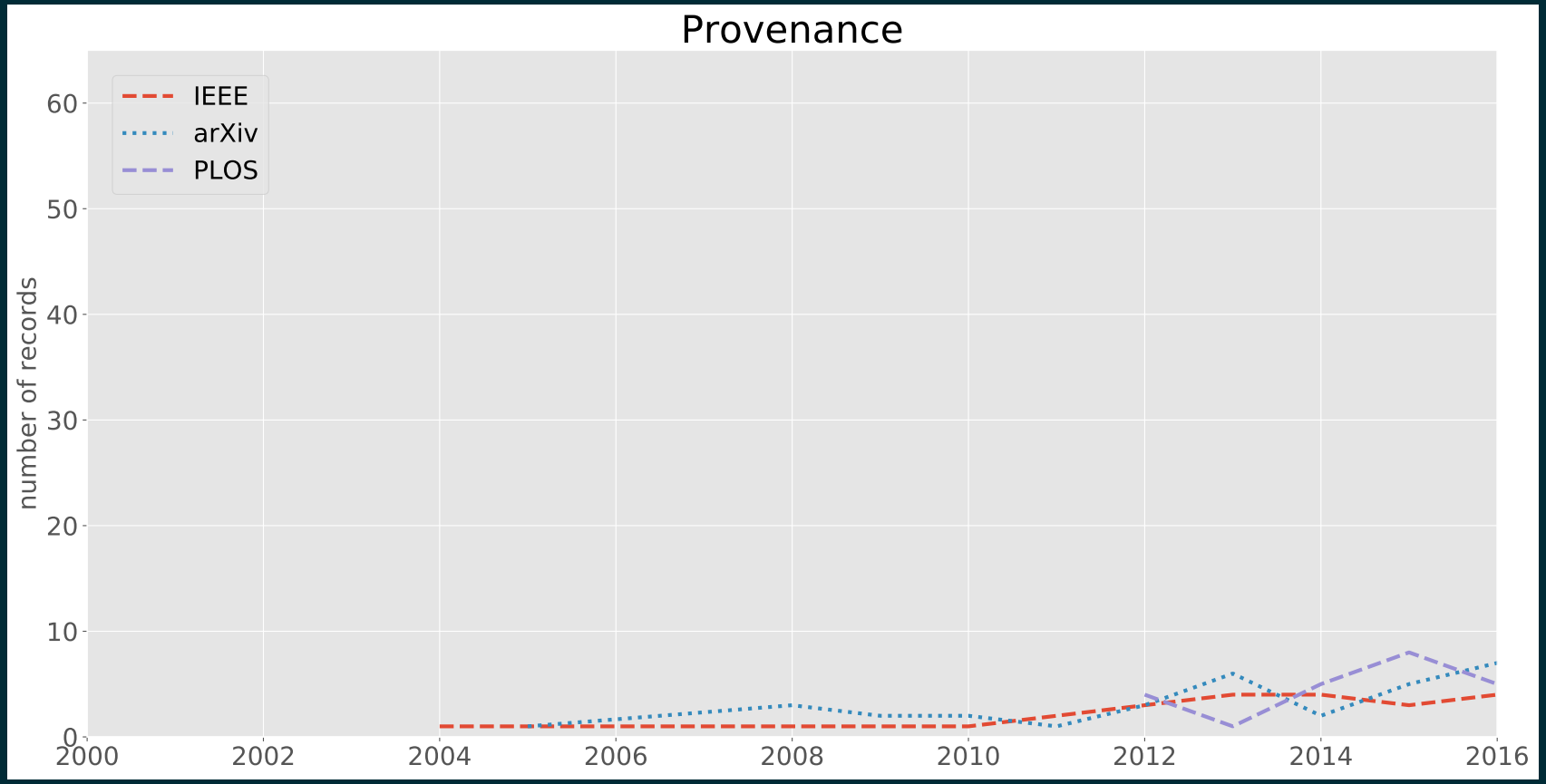
```
>>> import arcas

>>> api = arcas.Arxiv()
>>> parameters = api.parameters_fix(
...     title='sustainable software', records=1, start=1)
>>> url = api.create_url_search(parameters)
>>> request = api.make_request(url)
>>> root = api.get_root(request)
>>> raw_article = api.parse(root)

>>> article = api.to_dataframe(raw_article[0])
>>> api.export(article, "result.json")
```

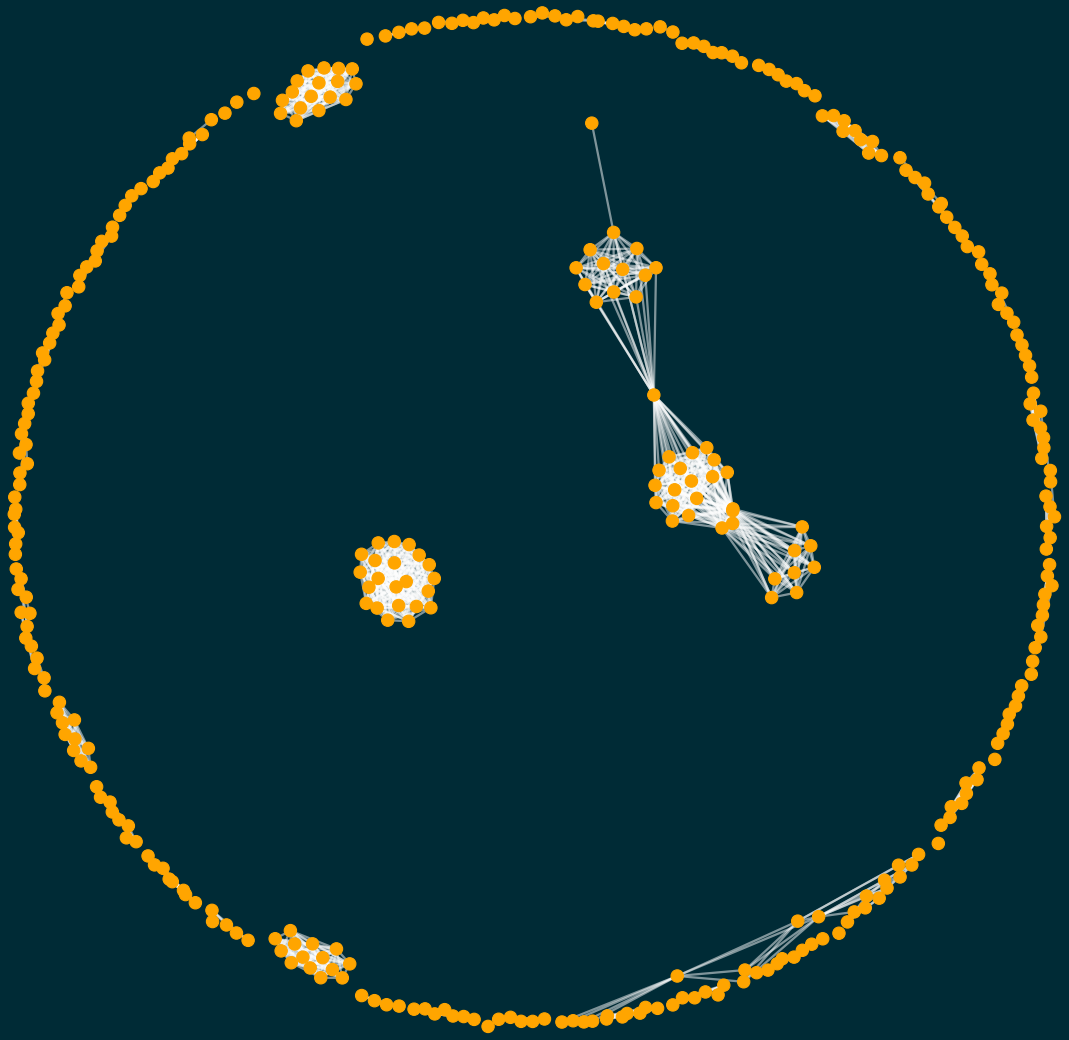
```
{"key":{"0":"Ahern2013"},
 "unique_key":{"0":"698d27415f69258ef122f46b184a77e0"},
 "title":{"0":"VisIt: Experiences with Sustainable Software"},
 "author":{"0":"Sean Ahern","1":"Eric Brugger"},
 "abstract":{"0":"  The success of the VisIt visualization..."},
 "date":{"0":2013},
 "journal":{"0":"arXiv"},
 "provenance":{"0":"arXiv"}}
```

```
>>> for p in [arcas.Arxiv, arcas.Nature, arcas.Ieee, arcas.Plos]:
...     api = p()
...     parameters = api.parameters_fix(
...          title='sustainable software', records=1, start=1)
...     url = api.create_url_search(parameters)
...     request = api.make_request(url)
...     root = api.get_root(request)
...     raw_article = api.parse(root)
...     try:
...         for art in raw_article:
...             article = api.to_dataframe(art)
...         api.export(article, "result_from_{}.json".format(
...             api.__class__.__name__))
...     except TypeError:
...         pass
```


Arcas
tools.py
doc/
arcas.readthedocs.io/
ieee
nature
arxiv
• • •
test_ieee
test_nature
test_arxiv
• • •


Articles per Year ($N = 87$)
number of records
16
14
12
10
8
6
4
2
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018


Provenance
IEEE
arXiv
PLOS
number of records
0
10
20
30
40
50
60
2000
2002
2004
2006
2008
2010
2012
2014
2016



https://github.com/ArcasProject/ArcasExamples

```
$ arcas_scrape --version
Arcas 0.0.1
```

```
$ arcas_scrape -p arxiv -t "Sustainable Software" -r 1
http://export.arxiv.org/api/query?search_query=ti:Sustainable Software&max_results=1&start=1
```

@NikoletaGlyn
https://github.com/Nikoleta-v3/Arcas
https://nikoleta-v3.github.io/talks/