Traceability, reproducibility, and scalability in Integrated Ecosystem Assessments:

June 2013

ECO-OP is supported by NSF Grant #0955649
PIs: Peter Fox (RPI) and Andrew Maffei (WHOI)
NEFSC Collaborators: Jon Hare and Mike Fogarty

Software programmer: Massimo Di Stefano
Informatics and metadata: Stace Beaulieu

Traceability, reproducibility, and scalability in Integrated Ecosystem Assessments: Adopting a provenance model for a collaborative report

June 2013

ECO-OP is supported by NSF Grant #0955649

PIs: Peter Fox (RPI) and Andrew Maffei (WHOI)
NEFSC Collaborators: Jon Hare and Mike Fogarty

Software programmer: Massimo Di Stefano
Informatics and metadata: Stace Beaulieu

Traceability, reproducibility, and scalability in Integrated Ecosystem Assessments: Adopting a provenance model for a collaborative report

June 2013

ECO-OP is supported by NSF Grant #0955649

PIs: Peter Fox (RPI) and Andrew Maffei (WHOI)
NEFSC Collaborators: Jon Hare and Mike Fogarty

Software programmer: Massimo Di Stefano
Informatics and metadata: Stace Beaulieu

Traceability, reproducibility, and scalability in Integrated Ecosystem Assessments: Adopting a provenance model for a collaborative report

June 2013

ECO-OP is supported by NSF Grant #0955649

PIs: Peter Fox (RPI) and Andrew Maffei (WHOI)
NEFSC Collaborators: Jon Hare and Mike Fogarty

Software programmer: Massimo Di Stefano
Informatics and metadata: Stace Beaulieu

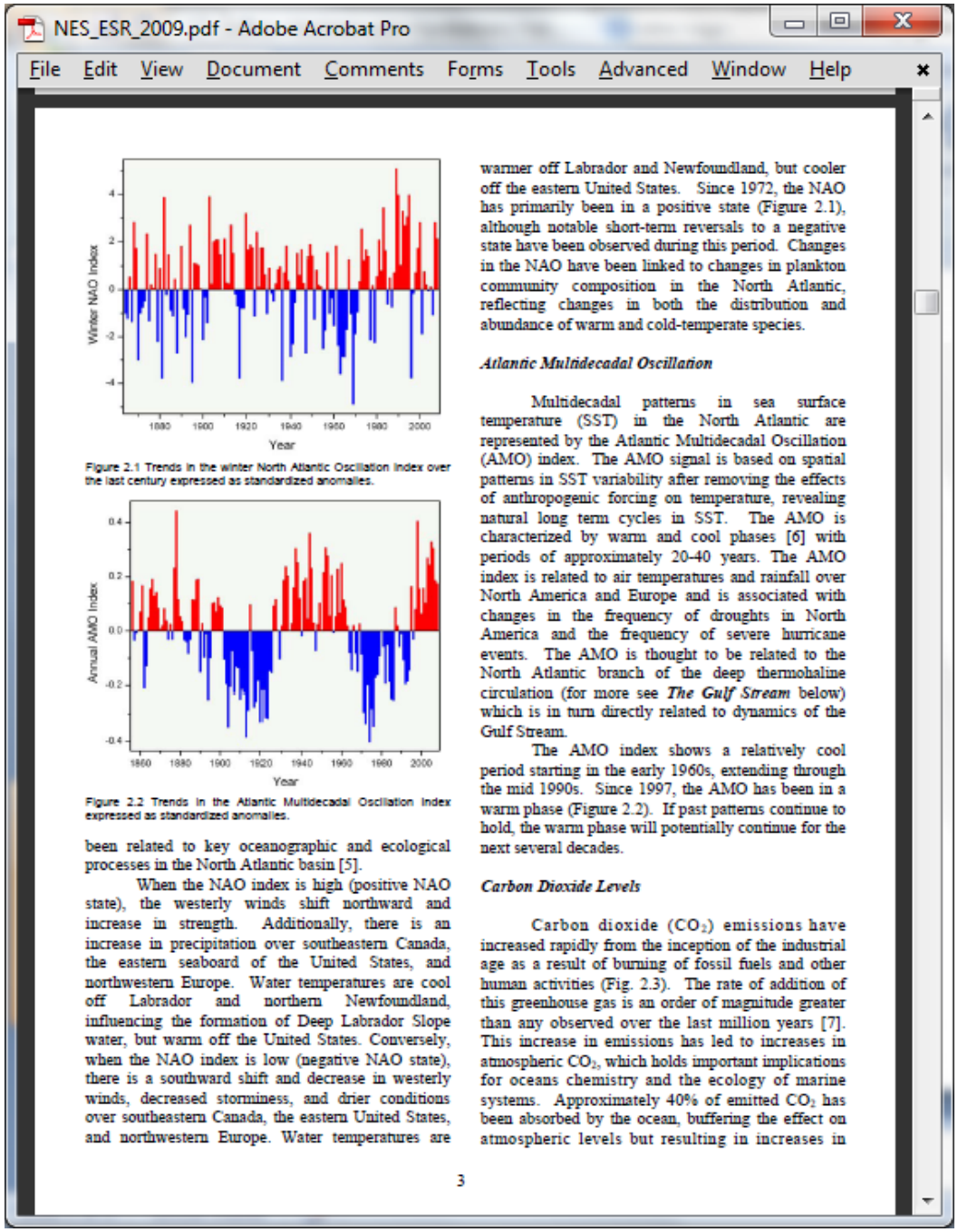
Provenance metadata

Use Case:

Northeast Shelf Large Marine Ecosystem Ecosystem Status Report

Goal:

"traceability, repeatability, explanation, verification, and validation" for ecosystem data and information products in the NEFSC Ecosystem Status Report (ESR)



NES_ESR_2009.pdf - Adobe Acrobat Pro

File Edit View Document Comments Forms Tools Advanced Window Help

Figure 2.1 Trends in the winter North Atlantic Oscillation index over the last century expressed as standardized anomalies.

Figure 2.2 Trends in the Atlantic Multidecadal Oscillation index expressed as standardized anomalies.

been related to key oceanographic and ecological processes in the North Atlantic basin [5].

When the NAO index is high (positive NAO state), the westerly winds shift northward and increase in strength. Additionally, there is an increase in precipitation over southeastern Canada, the eastern seaboard of the United States, and northwestern Europe. Water temperatures are cool off Labrador and northern Newfoundland, influencing the formation of Deep Labrador Slope water, but warm off the United States. Conversely, when the NAO index is low (negative NAO state), there is a southward shift and decrease in westerly winds, decreased storminess, and drier conditions over southeastern Canada, the eastern United States, and northwestern Europe. Water temperatures are warmer off Labrador and Newfoundland, but cooler off the eastern United States. Since 1972, the NAO has primarily been in a positive state (Figure 2.1), although notable short-term reversals to a negative state have been observed during this period. Changes in the NAO have been linked to changes in plankton community composition in the North Atlantic, reflecting changes in both the distribution and abundance of warm and cold-temperate species.

Atlantic Multidecadal Oscillation

Multidecadal patterns in sea surface temperature (SST) in the North Atlantic are represented by the Atlantic Multidecadal Oscillation (AMO) index. The AMO signal is based on spatial patterns in SST variability after removing the effects of anthropogenic forcing on temperature, revealing natural long term cycles in SST. The AMO is characterized by warm and cool phases [6] with periods of approximately 20-40 years. The AMO index is related to air temperatures and rainfall over North America and Europe and is associated with changes in the frequency of droughts in North America and the frequency of severe hurricane events. The AMO is thought to be related to the North Atlantic branch of the deep thermohaline circulation (for more see ***The Gulf Stream*** below) which is in turn directly related to dynamics of the Gulf Stream.

The AMO index shows a relatively cool period starting in the early 1960s, extending through the mid 1990s. Since 1997, the AMO has been in a warm phase (Figure 2.2). If past patterns continue to hold, the warm phase will potentially continue for the next several decades.

Carbon Dioxide Levels

Carbon dioxide (CO_2) emissions have increased rapidly from the inception of the industrial age as a result of burning of fossil fuels and other human activities (Fig. 2.3). The rate of addition of this greenhouse gas is an order of magnitude greater than any observed over the last million years [7]. This increase in emissions has led to increases in atmospheric CO_2, which holds important implications for oceans chemistry and the ecology of marine systems. Approximately 40% of emitted CO_2 has been absorbed by the ocean, buffering the effect on atmospheric levels but resulting in increases in

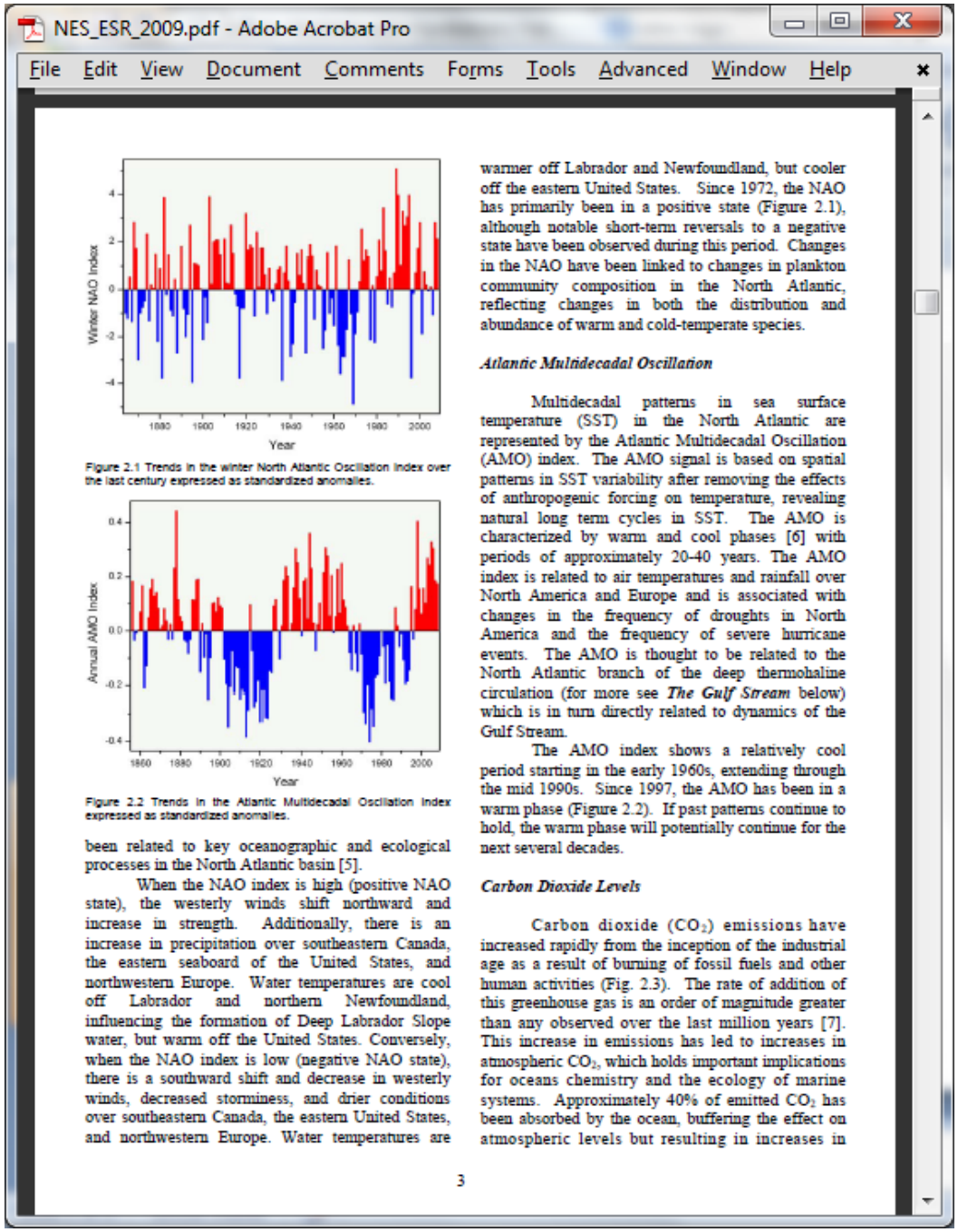
3

Page from 2009 ESR

Section on Climate Forcing

Figures available for download as PDF or image files –

<u>but without access to data or metadata</u>



NES_ESR_2009.pdf - Adobe Acrobat Pro

File Edit View Document Comments Forms Tools Advanced Window Help

Figure 2.1 Trends in the winter North Atlantic Oscillation index over the last century expressed as standardized anomalies.

Figure 2.2 Trends in the Atlantic Multidecadal Oscillation index expressed as standardized anomalies.

been related to key oceanographic and ecological processes in the North Atlantic basin [5].

When the NAO index is high (positive NAO state), the westerly winds shift northward and increase in strength. Additionally, there is an increase in precipitation over southeastern Canada, the eastern seaboard of the United States, and northwestern Europe. Water temperatures are cool off Labrador and northern Newfoundland, influencing the formation of Deep Labrador Slope water, but warm off the United States. Conversely, when the NAO index is low (negative NAO state), there is a southward shift and decrease in westerly winds, decreased storminess, and drier conditions over southeastern Canada, the eastern United States, and northwestern Europe. Water temperatures are warmer off Labrador and Newfoundland, but cooler off the eastern United States. Since 1972, the NAO has primarily been in a positive state (Figure 2.1), although notable short-term reversals to a negative state have been observed during this period. Changes in the NAO have been linked to changes in plankton community composition in the North Atlantic, reflecting changes in both the distribution and abundance of warm and cold-temperate species.

Atlantic Multidecadal Oscillation

Multidecadal patterns in sea surface temperature (SST) in the North Atlantic are represented by the Atlantic Multidecadal Oscillation (AMO) index. The AMO signal is based on spatial patterns in SST variability after removing the effects of anthropogenic forcing on temperature, revealing natural long term cycles in SST. The AMO is characterized by warm and cool phases [6] with periods of approximately 20-40 years. The AMO index is related to air temperatures and rainfall over North America and Europe and is associated with changes in the frequency of droughts in North America and the frequency of severe hurricane events. The AMO is thought to be related to the North Atlantic branch of the deep thermohaline circulation (for more see ***The Gulf Stream*** below) which is in turn directly related to dynamics of the Gulf Stream.

The AMO index shows a relatively cool period starting in the early 1960s, extending through the mid 1990s. Since 1997, the AMO has been in a warm phase (Figure 2.2). If past patterns continue to hold, the warm phase will potentially continue for the next several decades.

Carbon Dioxide Levels

Carbon dioxide (CO_2) emissions have increased rapidly from the inception of the industrial age as a result of burning of fossil fuels and other human activities (Fig. 2.3). The rate of addition of this greenhouse gas is an order of magnitude greater than any observed over the last million years [7]. This increase in emissions has led to increases in atmospheric CO_2, which holds important implications for oceans chemistry and the ecology of marine systems. Approximately 40% of emitted CO_2 has been absorbed by the ocean, buffering the effect on atmospheric levels but resulting in increases in

3

Page from 2009 ESR

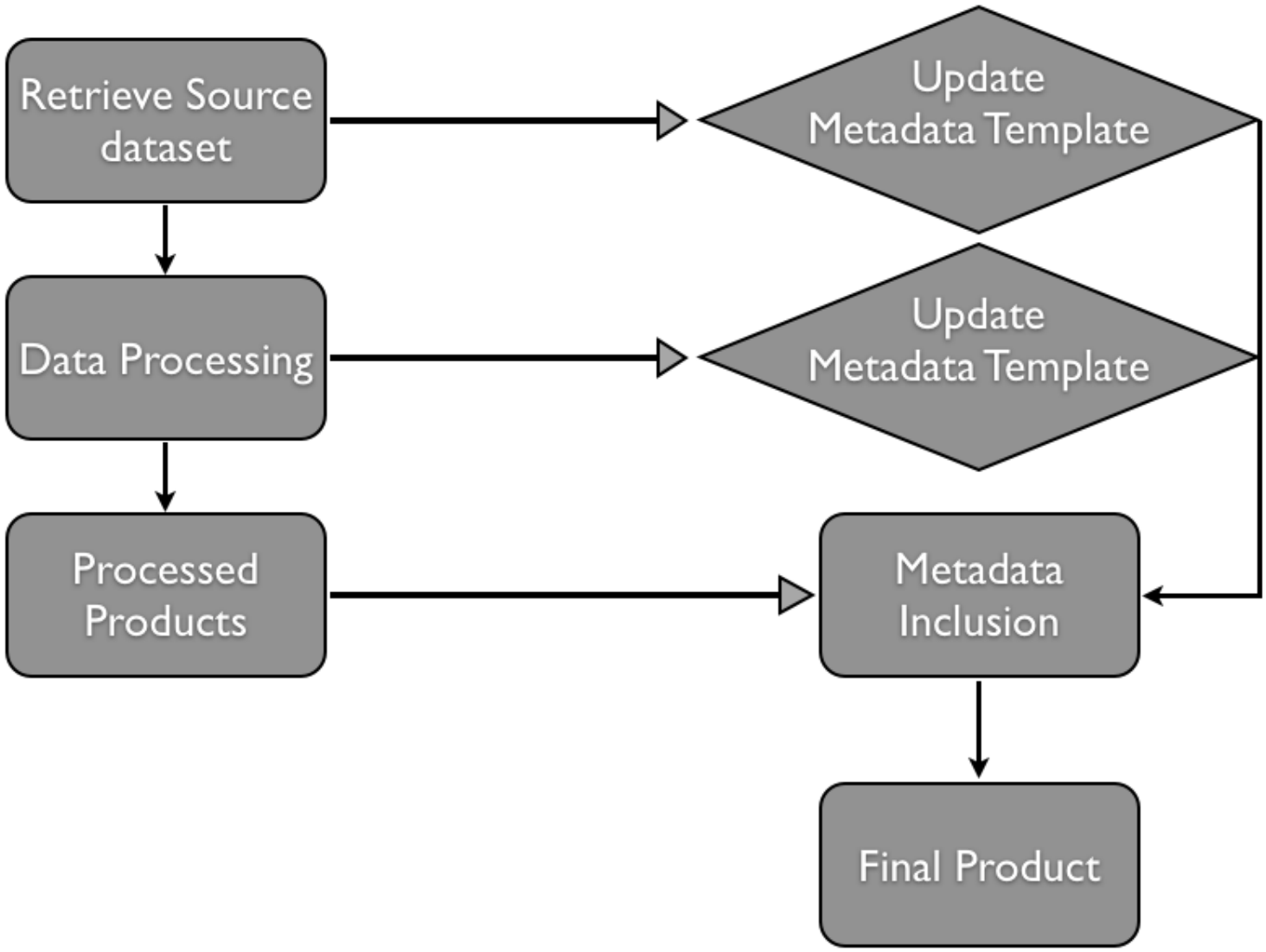
Section on Climate Forcing

Figures available for download as PDF or image files –

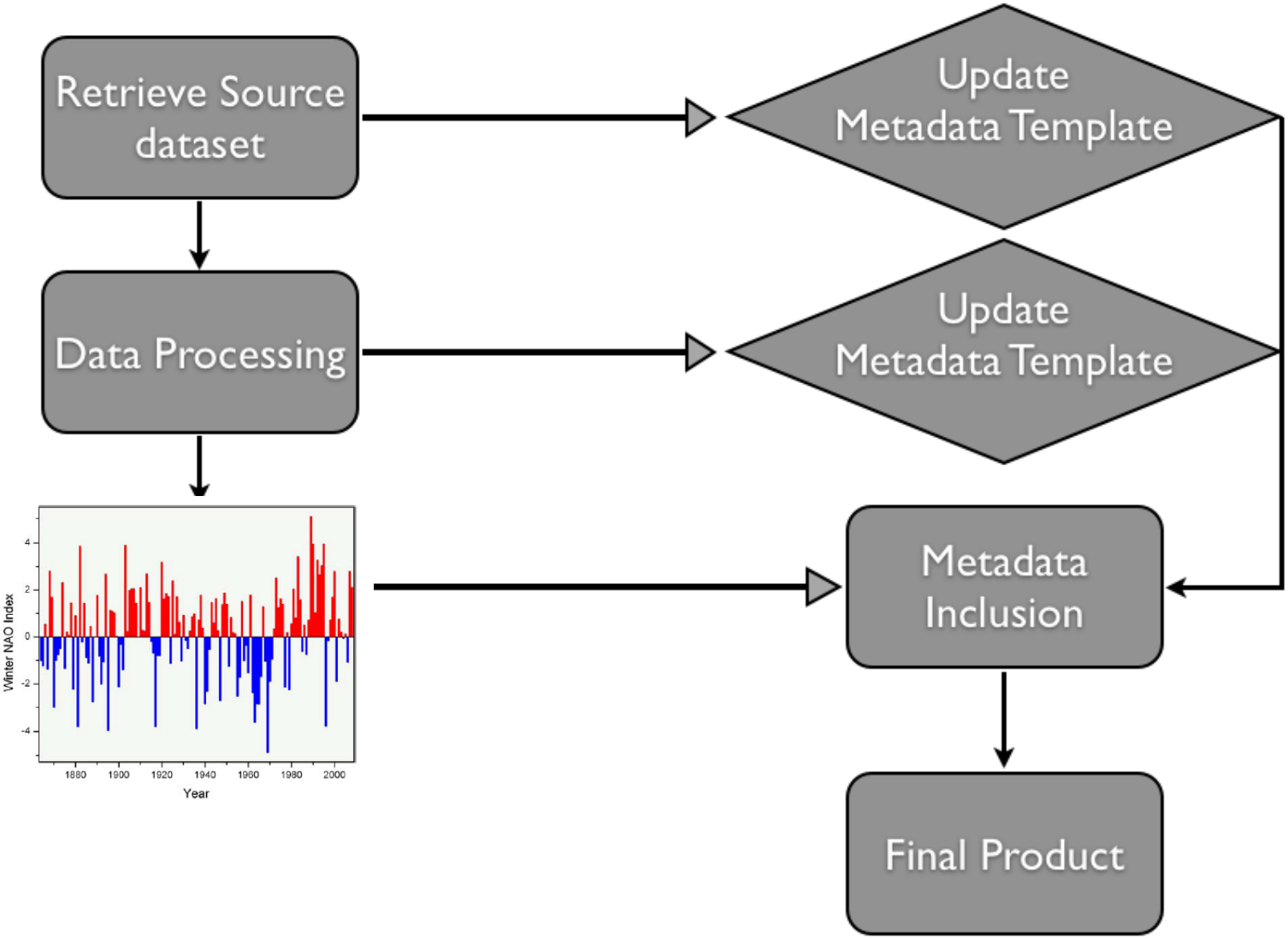
<u>but without access to data or metadata</u>

For the collaboration, we also want to track provenance metadata

Software design to track provenance



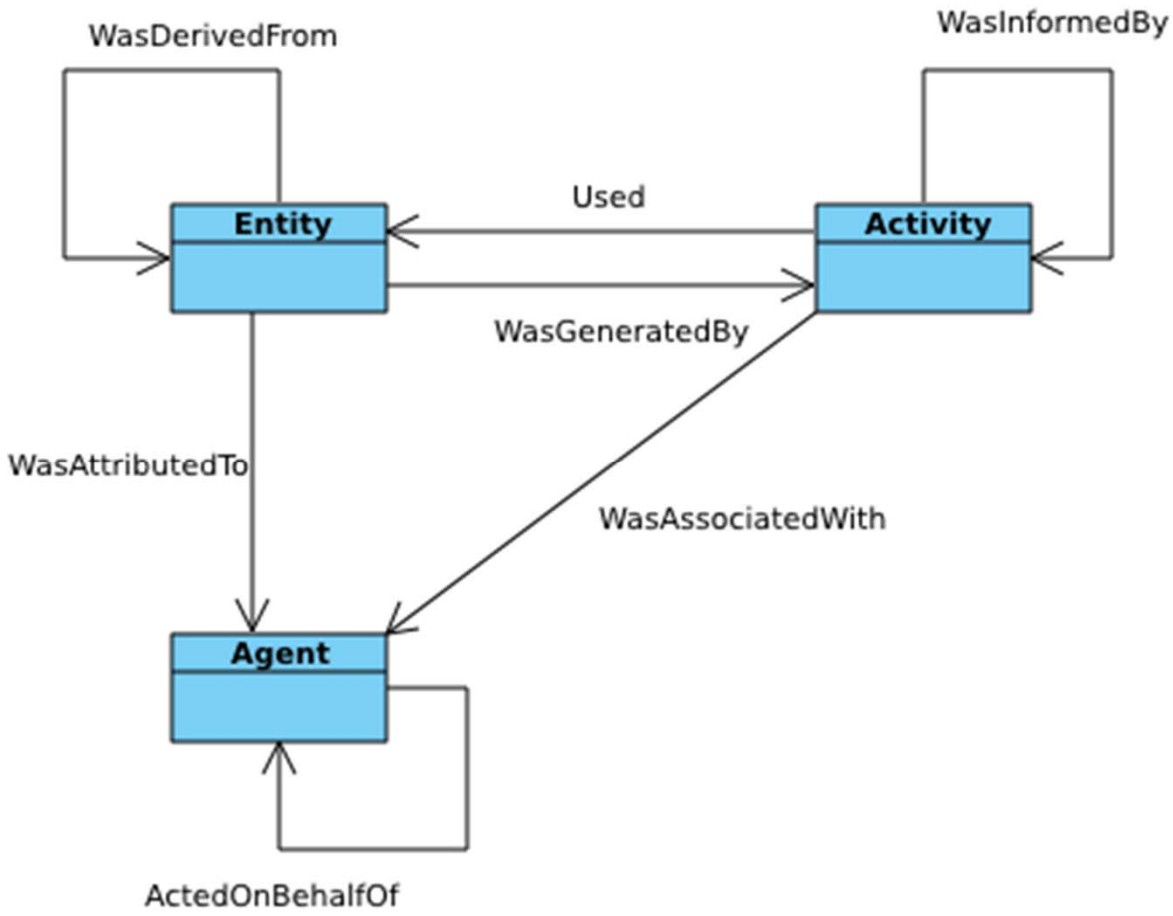
Software design to track provenance



PROV Data Model

http://www.w3.org/TR/prov-dm/
W3C Recommendation 30 April 2013

Core Structures (types and relations)

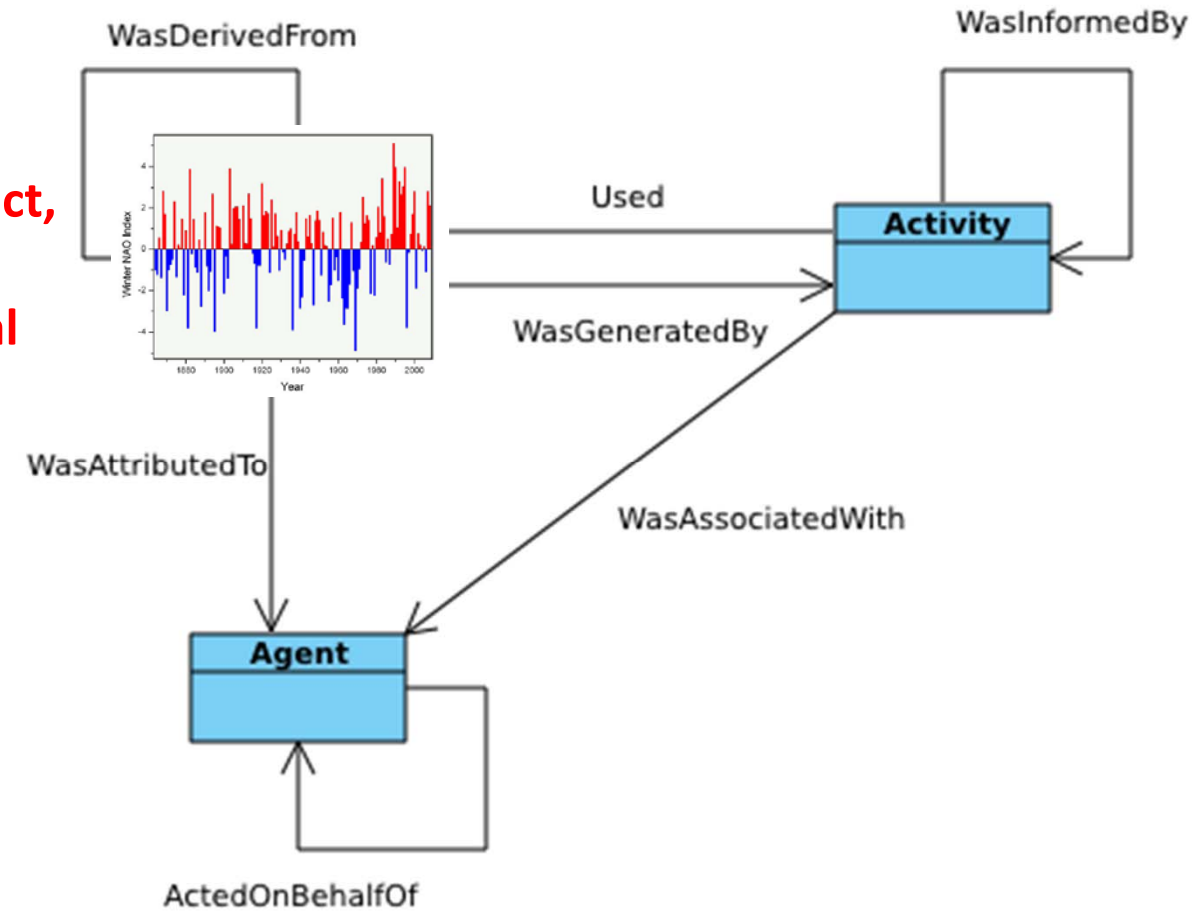


PROV Data Model

http://www.w3.org/TR/prov-dm/
W3C Recommendation 30 April 2013

Core Structures (types and relations)

Entity may be a single data product, or a chapter containing several data products

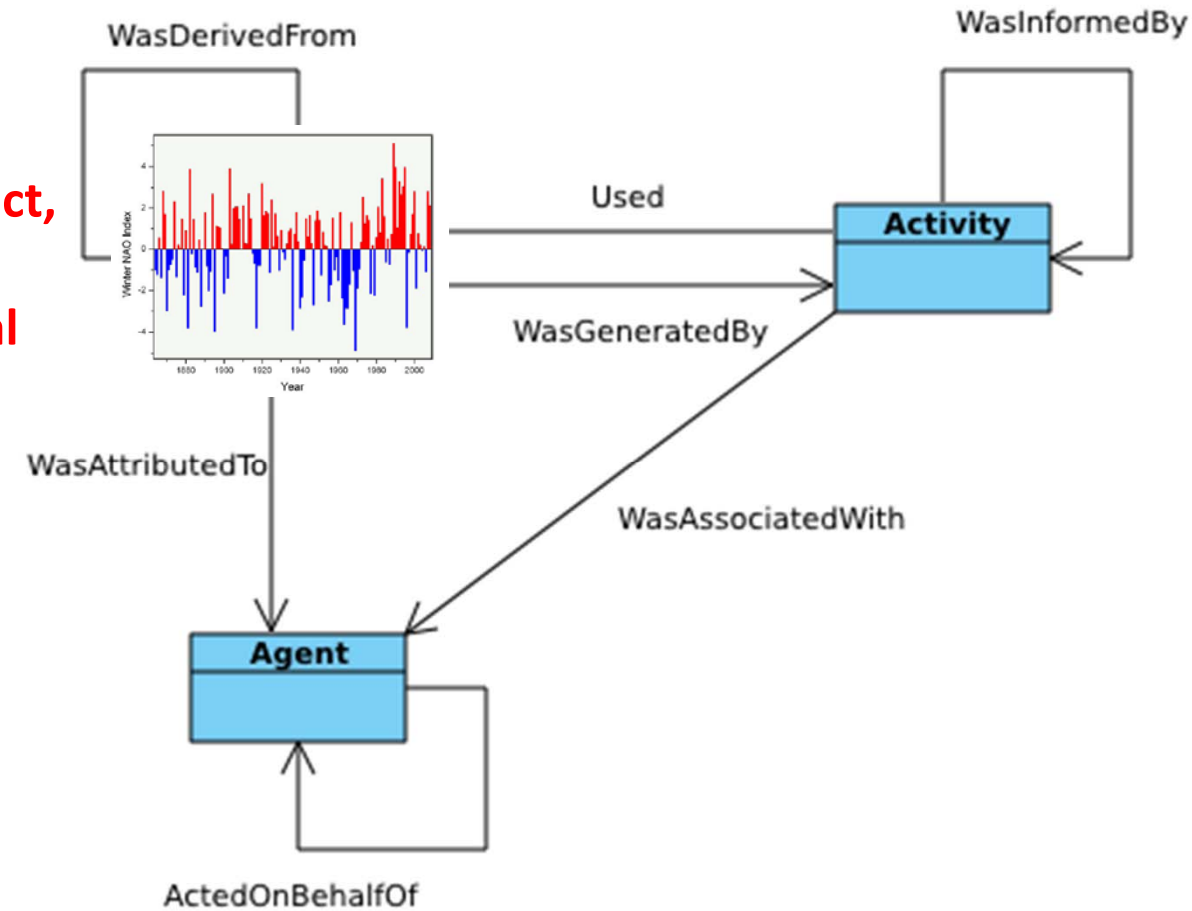


PROV Data Model

http://www.w3.org/TR/prov-dm/
W3C Recommendation 30 April 2013

Core Structures (types and relations)

Entity may be a single data product, or a chapter containing several data products



PROV-O: The PROV Ontology (expresses PROV-DM using OWL2)
http://www.w3.org/TR/prov-o/

Screenshot of IPython Notebook used to track both data and workflow provenance

http://ipython.org/

nao.csv | nbviewer.ipython.org

localhost:8888/033e19d0-ec1b-466b-aec5-24bf4f014973#

IP[y]: Notebook Ecosystem_Status_Report_w_metadata (autosaved) Logout

File Edit View Insert Cell Kernel Help

Code Cell Toolbar: None

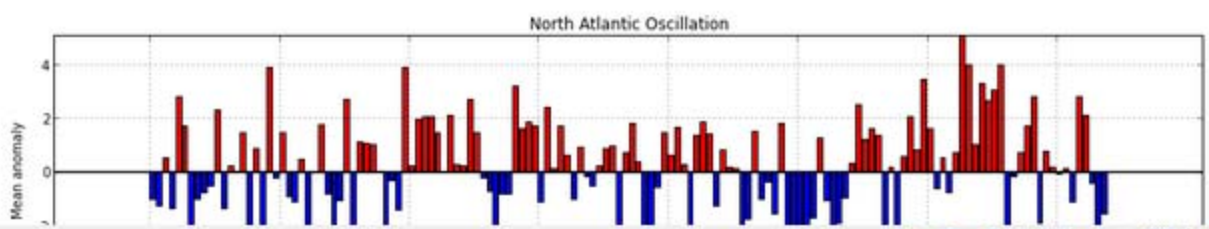
Terminal

Acquire, process, and plot NAO and AMO

In [3]:
```
nao = d.getnao(output=ID, savetype='csv')
NAO['execution_time'] = datetime.datetime.now()
amo = d.getamo(output=ID, savetype='csv')
AMO['execution_time'] = datetime.datetime.now()
d.plotDriver(nao, ptype='bar', name='nao', nb='y', xticks=20, xticks
             figsize=(6,4), xmargin=True, ymargin=True, legend=Fals
             grid=True, ylabel='Mean anomaly', title='North Atlanti

d.plotDriver(amo, ptype='bar',name='amo', nb='y', xticks=20, xticks_
             figsize=(6,4), xmargin=True, ymargin=True, legend=Fals
             grid=True, ylabel='Mean anomaly', title='Atlantic Mult
```

```
NAO data saved in :  epitest_Wednesday_19_June_2013_04_03_09_PM
AMO data saved in :  epitest_Wednesday_19_June_2013_04_03_09_PM
graph saved in:  epitest_Wednesday_19_June_2013_04_03_09_PM/nao_bar
.png
```



Right Ctrl

Screenshot of IPython Notebook used to track both data and workflow provenance

http://ipython.org/

nao.csv | nbviewer.ipython.org

localhost:8888/033e19d0-ec1b-466b-aec5-24bf4f014973#

IP[y]: Notebook Ecosystem_Status_Report_w_metadata (autosaved) Logout

File Edit View Insert Cell Kernel Help

Code Cell Toolbar: None

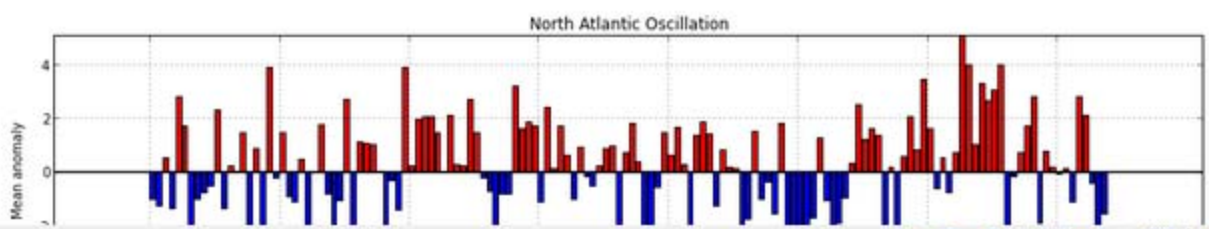
Terminal

Acquire, process, and plot NAO and AMO

In [3]:
```
nao = d.getnao(output=ID, savetype='csv')
NAO['execution_time'] = datetime.datetime.now()
amo = d.getamo(output=ID, savetype='csv')
AMO['execution_time'] = datetime.datetime.now()
d.plotDriver(nao, ptype='bar', name='nao', nb='y', xticks=20, xticks
             figsize=(6,4), xmargin=True, ymargin=True, legend=Fals
             grid=True, ylabel='Mean anomaly', title='North Atlanti

d.plotDriver(amo, ptype='bar',name='amo', nb='y', xticks=20, xticks_
             figsize=(6,4), xmargin=True, ymargin=True, legend=Fals
             grid=True, ylabel='Mean anomaly', title='Atlantic Mult
```

```
NAO data saved in :  epitest_Wednesday_19_June_2013_04_03_09_PM
AMO data saved in :  epitest_Wednesday_19_June_2013_04_03_09_PM
graph saved in:  epitest_Wednesday_19_June_2013_04_03_09_PM/nao_bar
.png
```



Right Ctrl

Code in Python, Matlab, R, other

Screenshot of IPython Notebook used to track both data and workflow provenance

http://ipython.org/

nao.csv | nbviewer.ipython.org

localhost:8888/033e19d0-ec1b-466b-aec5-24bf4f014973#

IP[y]: Notebook Ecosystem_Status_Report_w_metadata (autosaved) Logout

File Edit View Insert Cell Kernel Help

Code Cell Toolbar: None

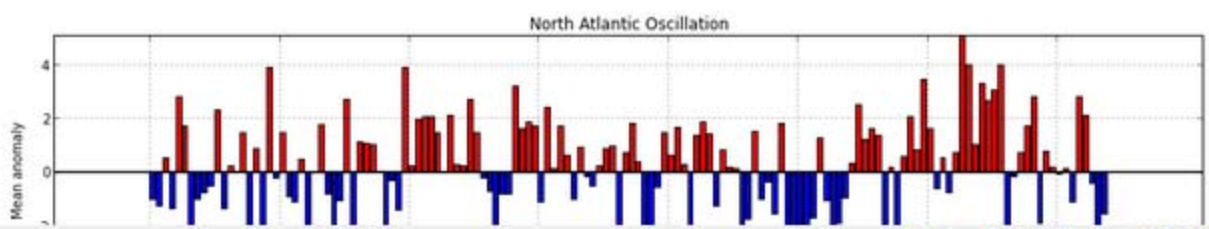
Terminal

Acquire, process, and plot NAO and AMO

In [3]:
```
nao = d.getnao(output=ID, savetype='csv')
NAO['execution_time'] = datetime.datetime.now()
amo = d.getamo(output=ID, savetype='csv')
AMO['execution_time'] = datetime.datetime.now()
d.plotDriver(nao, ptype='bar', name='nao', nb='y', xticks=20, xticks
             figsize=(6,4), xmargin=True, ymargin=True, legend=Fals
             grid=True, ylabel='Mean anomaly', title='North Atlanti

d.plotDriver(amo, ptype='bar',name='amo', nb='y', xticks=20, xticks_
             figsize=(6,4), xmargin=True, ymargin=True, legend=Fals
             grid=True, ylabel='Mean anomaly', title='Atlantic Mult
```

```
NAO data saved in :  epitest_Wednesday_19_June_2013_04_03_09_PM
AMO data saved in :  epitest_Wednesday_19_June_2013_04_03_09_PM
graph saved in:  epitest_Wednesday_19_June_2013_04_03_09_PM/nao_bar
.png
```



Right Ctrl

Code in Python, Matlab, R, other

Screenshot of IPython Notebook used to track both data and workflow provenance

http://ipython.org/

nao.csv | nbviewer.ipython.org

localhost:8888/033e19d0-ec1b-466b-aec5-24bf4f014973#

IP[y]: Notebook Ecosystem_Status_Report_w_metadata (autosaved) Logout

File Edit View Insert Cell Kernel Help

Code Cell Toolbar: None

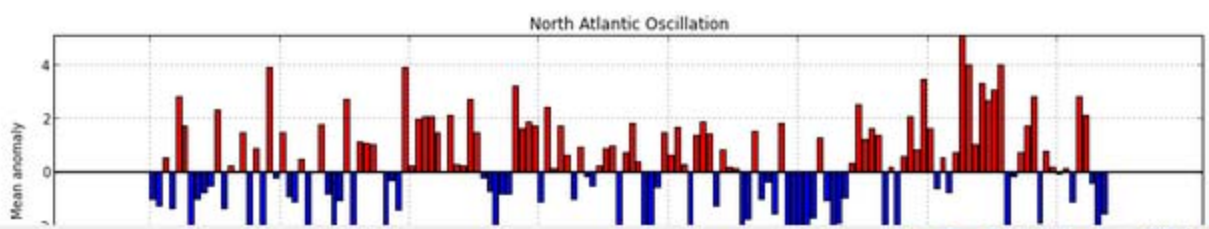
Terminal

Acquire, process, and plot NAO and AMO

In [3]:
```
nao = d.getnao(output=ID, savetype='csv')
NAO['execution_time'] = datetime.datetime.now()
amo = d.getamo(output=ID, savetype='csv')
AMO['execution_time'] = datetime.datetime.now()
d.plotDriver(nao, ptype='bar', name='nao', nb='y', xticks=20, xticks
             figsize=(6,4), xmargin=True, ymargin=True, legend=Fals
             grid=True, ylabel='Mean anomaly', title='North Atlanti

d.plotDriver(amo, ptype='bar',name='amo', nb='y', xticks=20, xticks_
             figsize=(6,4), xmargin=True, ymargin=True, legend=Fals
             grid=True, ylabel='Mean anomaly', title='Atlantic Mult
```

```
NAO data saved in :  epitest_Wednesday_19_June_2013_04_03_09_PM
AMO data saved in :  epitest_Wednesday_19_June_2013_04_03_09_PM
graph saved in:  epitest_Wednesday_19_June_2013_04_03_09_PM/nao_bar
.png
```



Right Ctrl

Notebook can be shared, or output as script, HTML, PDF, other

PDF output of IPython Notebook with clickable links to data and code

lantic, components of the large-scale circulation of the Atlantic Ocean, and issues related to ocean acidification.

1.1 North Atlantic Oscillation Index

Climate and weather over the North Atlantic are strongly influenced by the relative strengths of two large-scale atmospheric pressure cells – the Icelandic Low and the Azores High [4]. As the relative strengths of these two pressure systems vary, characteristic patterns of temperature, precipitation, and wind fields are observed. An index of this dipole pattern has been developed based on the standardized difference in sea level pressure between Lisbon, Portugal and Reykjavík, Iceland in the winter (December-February; see Glossary for a description of methods used to create standardized indicators). This North Atlantic Oscillation (NAO) index has been related to key oceano-

eastern United States. Since 1972, the NAO has primarily been in a positive state (Figure 1), although notable short-term reversals to a negative state have been observed during this period. Changes in the NAO have been linked to changes in plankton community composition in the North Atlantic, reflecting changes in both the distribution and abundance of warm and cold-temperate species.

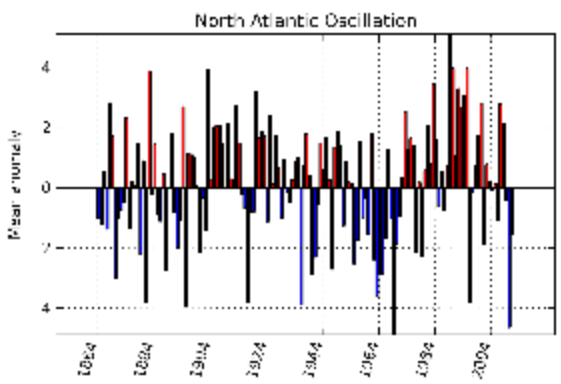


Figure 1: North Atlantic Oscillation - data - code

PDF output of IPython Notebook with clickable links to data and code

lantic, components of the large-scale circulation of the Atlantic Ocean, and issues related to ocean acidification.

1.1 North Atlantic Oscillation Index

Climate and weather over the North Atlantic are strongly influenced by the relative strengths of two large-scale atmospheric pressure cells – the Icelandic Low and the Azores High [4]. As the relative strengths of these two pressure systems vary, characteristic patterns of temperature, precipitation, and wind fields are observed. An index of this dipole pattern has been developed based on the standardized difference in sea level pressure between Lisbon, Portugal and Reykjavík, Iceland in the winter (December-February; see Glossary for a description of methods used to create standardized indicators). This North Atlantic Oscillation (NAO) index has been related to key oceano-

eastern United States. Since 1972, the NAO has primarily been in a positive state (Figure 1), although notable short-term reversals to a negative state have been observed during this period. Changes in the NAO have been linked to changes in plankton community composition in the North Atlantic, reflecting changes in both the distribution and abundance of warm and cold-temperate species.

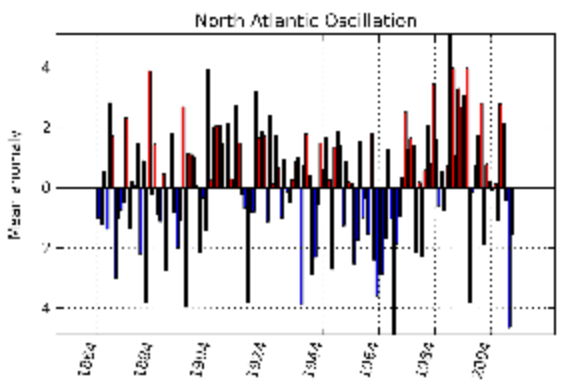


Figure 1: North Atlantic Oscillation - data - code

Screenshot of csv file at GitHub

IPy IPython Dashboard | IPy Ecosystem_Status_Repor | IPy Ecosystem_Status_Repor | nao.csv | nbvie

https://gist.github.com/anonymous/5815519

GitHub Gist

Sign up for a Git

Chromium Web Browser / nao.csv

Created 3 hours ago

Gist Detail

Revisions 1

Download Gist

Clone this gist

/anonymous/5815519

Embed this gist

<script src="https:/

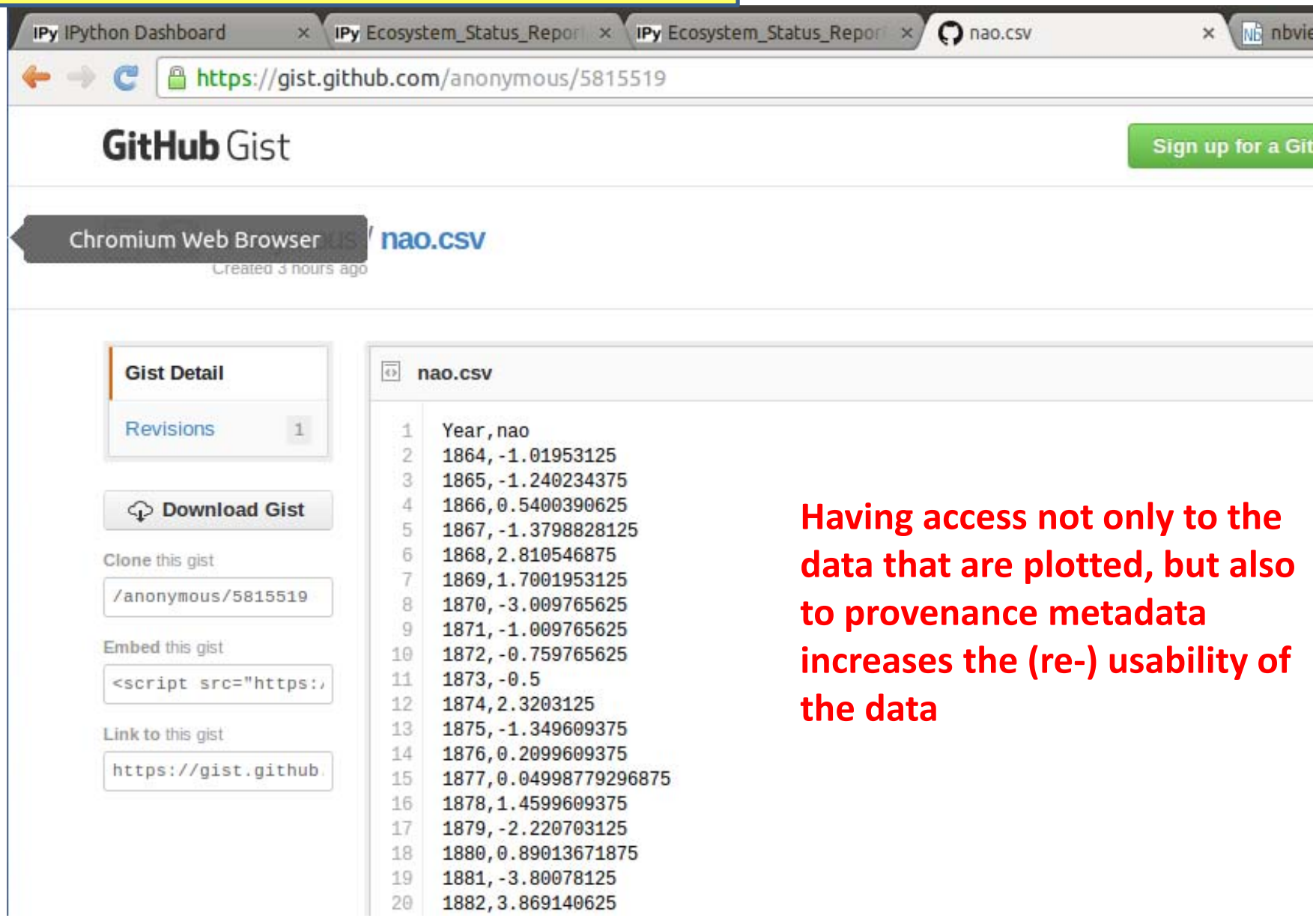
Link to this gist

https://gist.github

nao.csv

```
Year,nao
1864,-1.01953125
1865,-1.240234375
1866,0.5400390625
1867,-1.3798828125
1868,2.810546875
1869,1.7001953125
1870,-3.009765625
1871,-1.009765625
1872,-0.759765625
1873,-0.5
1874,2.3203125
1875,-1.349609375
1876,0.2099609375
1877,0.04998779296875
1878,1.4599609375
1879,-2.220703125
1880,0.89013671875
1881,-3.80078125
1882,3.869140625
```

Screenshot of csv file at GitHub



Having access not only to the data that are plotted, but also to provenance metadata increases the (re-) usability of the data