



Preservation Tools Technologies and Policies Interest Group PTTP-IG

Breakout-6: Thursday, 12th November 2020 00:00-01:30 UTC

Co-chairs: Ruth Duerr, Mike Hildreth, Peter Cornwell

Researchers' Tools for Knowledge Preservation

Agenda

- Long-term preservation strategies - AARNet update (Adam Bell, AARNet), Invenio RDM launch update (Lars Holm, CERN) and OCFL launch update (Neil Jefferies, Oxford)
- FAIR update (Mike Hildreth, PTTP), DataAtRisk.org (Ruth Duerr, PTTP)
- Annotation productivity tools for COVID conditions (Peter Cornwell, PTTP; Eric Decker, Basel)

Reference Materials

This meeting provides a forum for discussion about Knowledge Preservation, and in particular key technologies and tools for researchers. We will focus on recent work that broadens the scope and applicability of the preservation of scientific annotation, hearing from several projects that have not yet received attention in this forum.

Preliminary version of slides available at <https://www.data-futures.org/pages/events>

Please use the collaborative notes document : <https://docs.google.com/document/d/1Yan9zII-7a3Y75mzmqRfHFcnphPoWya8ry6WdsV6TuA/edit> for discussion proposals.

Annotation productivity tools for COVID working conditions

- during Spring 2020 and again in October/November many organizations have been forced to re-structure student and staff working practices as well as research agendas
- Basel (Switzerland), ENS-Lyon (France) and Notre Dame (Indiana) agreed to share newly remote workforces and research materials, and have collaborated on research tasks that would otherwise be halted

cost-neutral collaboration produced new research data resources

- annotation infrastructure already in use by Basel, ENS and Notre Dame's Hesburgh Libraries was upgraded using CERN's Invenio OAuth server and ORCID sign-in, providing secure managed workflows
- work-package documentation and contributor management, as well as research tasks and personnel were provided by Hesburgh; research tasks and personnel by ENS, and research tasks and analysis and preservation in Invenio-based *hasdai* repositories undertaken by Basel

Organizing Cross-Institute Annotation Work during Lock-down

Don Brower – University of Notre Dame

Quick-start international collaboration

- *hasdai* partnership with CERN and the Hesburgh Libraries organized a remote, and previously-unscheduled annotation work-plan during the Spring Semester 2020
- ad-hoc international workforce: undergraduate students, some graduate students, plus some staff in France, Germany, Switzerland and U.S.
- included many people who were otherwise unable to work due to COVID stay-at-home orders

Secure annotation infrastructure

- ORCID sign-in to identify workers and provide secure access to workflow <https://auth.hasdai.org/>
- contributors not having an ORCID had to register one; whitelist for team management and rejection of unwanted contributors
- annæstore cloud annotation service divided each annotation task into discrete work-units, tracked progress and enabled contributors to review and re-work units they'd already completed

Annotations Projects—Remote Work Information

Getting Started

Project Instructions

[Directories & Chronicles](#)[Scanlan Map Instructions](#)[Irish Ballad Instructions](#)[Epistemological Letters
Phase II \(COMPLETE\)](#)[Vietnam-Information Booklet
Instructions \(COMPLETE\)](#)

Project Outcomes & Stories

Annotation Questions



Don Brower

DIRECTORIES & CHRONICLES

Project summary: The Asian Directories & Chronicles serial was published annually by The Hong Kong Daily Press between 1863 and 1941. Each volume provides listings of active corporations, foreign residents and government agencies of all nationalities for that year, together with their addresses in countries including Borneo, China, Indo-China, Japan, Korea and the Philippines. With the current exceptions of 1866, 1867, 1872, 1875 and 1884, all of the volumes of the Directories & Chronicles have been assembled by the Europa Institute at the University of Basel. High-resolution digitization of all the pages of the volumes and subsequent OCR and analysis has been used to create a **IIIF service**. Both manual and automated annotation of the pages has already identified sections of each of the volumes—creating IIIF manifests for the foreign resident and company listings and treaties. This project will commence analysis of the treaty material—initially identifying the titles of treaties and agreements, since the original publications do not contain tables of contents.

The first task is to identify the starting page of each treaty.

- [Instructions](#)
- [Starting Portal](#)

FAQ

Q: Should small caps be transcribed as lower case or upper case?

A: As lower case.

Q: Should headings with footnotes include the footnote text somehow?

A: Not at this point. Transcribe the asterisk but not the footnote.

Q: Are line breaks needed between a heading and a subheading?

A: Yes. Transcribe horizontal delimiters as a line break.

Reduced training needs

- administration website and shared how-to googledocs
- same sign-in and work-on 'task' how-to document for all annotation tasks
- detailed task-specific instruction documents and 'helpdesk' via email, with possibility of video-call follow-up
- no other training provided

Work on the East Asian Treaties task

Work on the Scanlan Maps task

Work on the Irish Ballads task

Work on the Epistemological Letters task (complete)

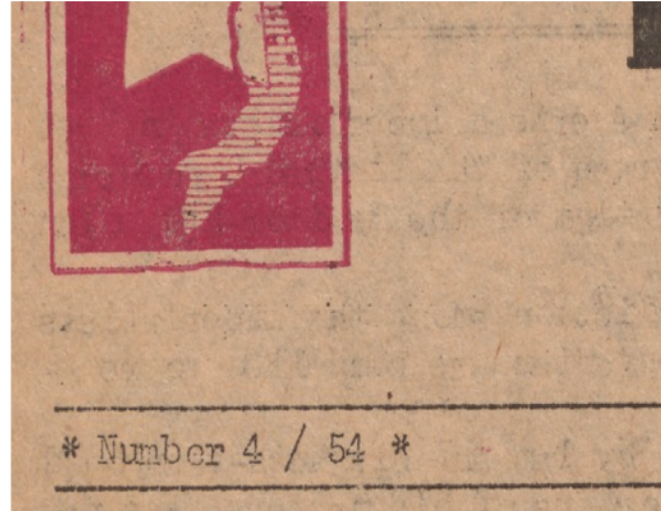
Work on the Vietnam Information task (complete)

preservable research outputs

- annotation collections produced against IIF services form basis of standards-based WADM datasets—can be consumed by multiple applications and are accepted by repositories such as Zenodo e.g. <https://doi.org/10.5281/zenodo.2580998>
- two Hesburgh annotation tasks have already produced independent Invenio repositories—including digitized sources as well as annotation collections forming primary research assets
 - Vietnam-Information: <https://vni.ens-lyon.hasdai.org/>
 - Asian Directories Treaties: <https://adc-treaties.unibas.hasdai.org/>



Search by Year



Search by Issue



Tables of Contents

Vietnam-Information Newsletter Corpus

The Vietnam-Information newsletters were a series of publications directed at the international community made by the Democratic Republic of Vietnam from the 1940s to the mid 1950s. The rare english-language booklets are each around 10 pages long and the Institute of East Asian Studies at ENS-Lyon (IAO) has the largest surviving collection, though the originals are now very fragile.

Maison des Sciences de l'Homme support has enabled hi-resolution digitization of the newsletters, preserving them and making them accessible to scholars. They are delivered here via an Invenio repository which supports research workflows using WADM annotation. Annotation already conducted in a collaboration with Hesburgh Libraries, Notre Dame, Indiana, has already identified issue metadata and transcribed tables of contents. As a result, the corpus can be searched by year and issue number, and work in progress will use the ToC data, which can currently be viewed in Mirador, to implement searching by article and subject.

Access the *hasdai* repository to browse the Vietnam-Information Newsletter Corpus

category

- ☐ Geography (164)
- ☐ Event (137)
- ☐ Persons (103)
- ☐ Institution (97)
- ☐ News (69)

locations

- ☐ Indochina (164)
- ☐ Vietnam (156)
- ☐ France (89)
- ☐ Laos (88)
- ☐ Cambodia (52)
- ☐ US (43)
- ☐ Switzerland (34)
- ☐ USSR (33)
- ☐ China (19)
- ☐ Korea (7)

persons

found 173 results

< 1 2 3 4 5 6 7 8 9 >



Viet-Nam Bulletin

date: 1956-10-22

issue: 31

lead article: 1 - Ha-noi After Two Years of Liberation



Viet-Nam Bulletin

date: 1956-10-12

issue: 30

lead article: 1 - Vietnamese Parliamentary Delegation to Visit the Soviet Union

East-Asian Treaties annotation task



Detecting and analysing multiple instances of the same document

Y OF FRIENDSHIP AND COMMERCE

fications Exchanged at Bangkok, 15th April, 1856

TREATY OF FRIENDSHIP AND COMMERCE BETWEEN
HER MAJESTY THE QUEEN OF THE UNITED KINGDOM AND
THE KINGS OF SIAM.

TREATY (_____)

Ratificati

Y OF FRIENDSHIP AND COMMERCE BETWEEN HER MAJESTY
THE QUEEN OF THE UNITED KINGDOM AND THE
KINGS OF SIAM.

Ratifications Exchanged at Bangkok, 5th April, 1856.

TREATY OF FRIENDSHIP AND COMMERCE BETWEEN
HER MAJESTY THE QUEEN OF THE UNITED KINGDOM
AND THE KINGS OF SIAM

Ratifications Exchanged at Bangkok, 15th April, 1856

TREATY OF FR
THE Q

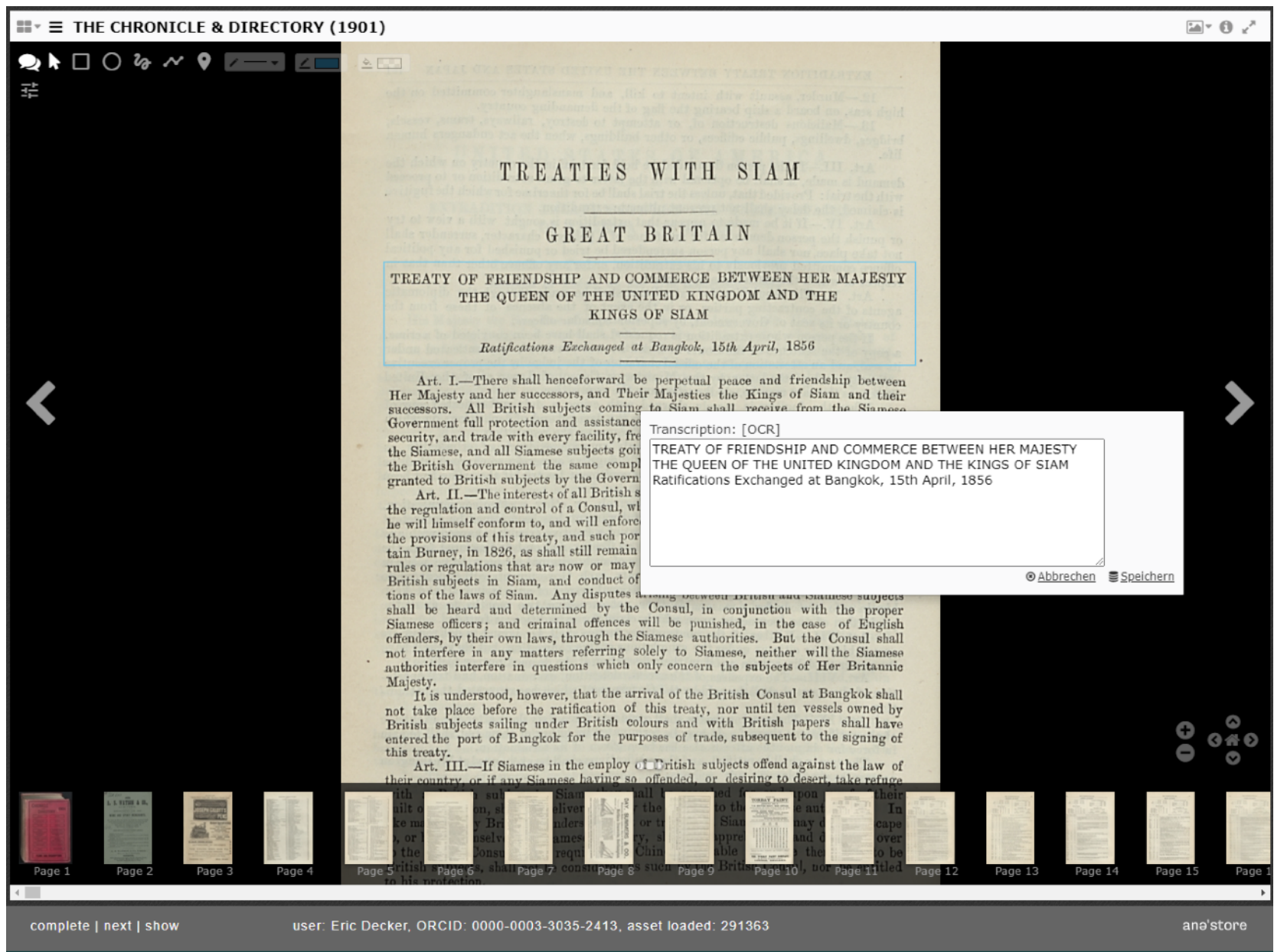
Ra

Manual annotation with interactive OCR support

In a three phases approach the aims were to (1) **identify page ranges** of the treaties sections in the corpus (2) **identify titles of documents** in these section only (3) **normalize** and classify titles.

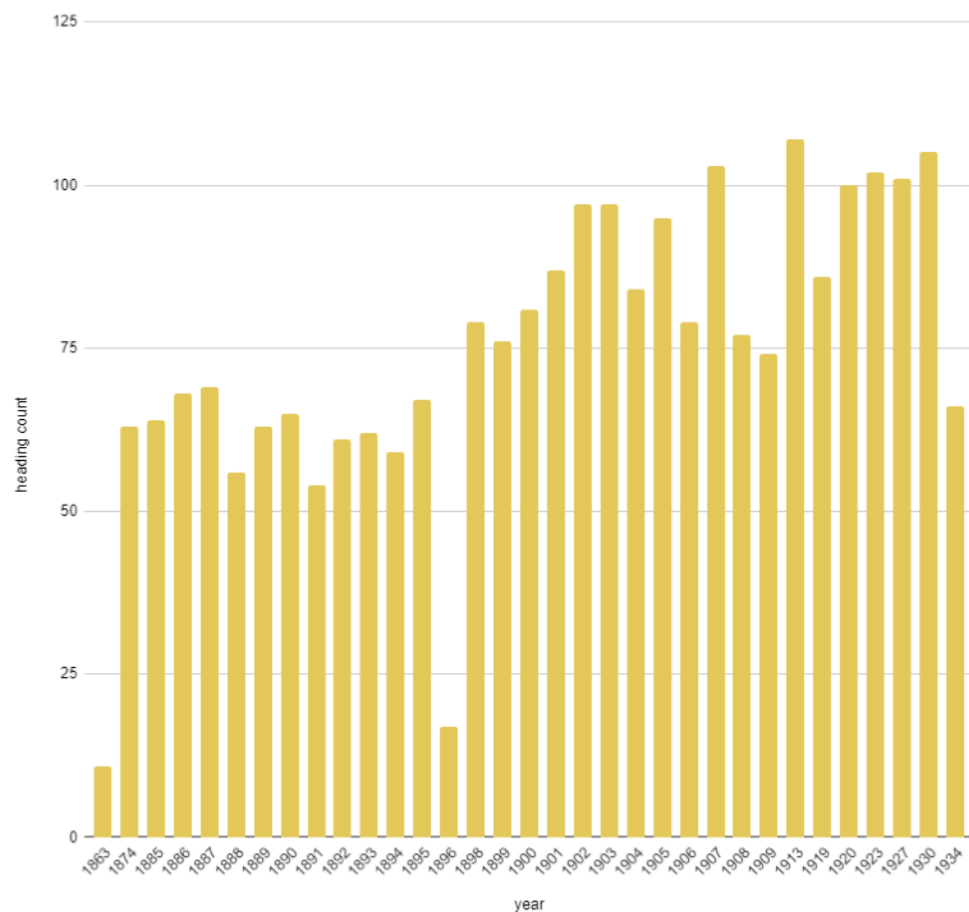
In phase two headers of treaty documents were annotated in a sample data set covering 33 years between 1863 and 1934.

A **set of criteria** for relevant headings was defined by a curator. A team at Notre Dame then annotated the **headings** and corrected the **OCR**.



Refining annotation output

document headings per year



In that process **2400+ headings** were detected and turned into **annotation assets**. Due to errors in print and OCR as well as changing naming conventions in the corpus and segmentation decisions made by annotators the **degree of variations** in titles describing the same document is **high**.

Up to 27 variations of the same document title were found


Annotation assets

Identifier	Transcribed Text	iiif Fragment	Annotator (ORCID)
https://iiif.directories.europa-institute.org/image/590037/canvas/p527/a0	SUPPLEMENTARY TREATY BETWEEN FRANCE AND JAPAN, SIGNED AT PARIS ON THE 20TH JUNE, 1864.	https://iiif.directories.europa-institute.org/image/594037_0000/45,746,1509,186/pct:33/0/default.jpg	0000-0003-3035-2413


Data normalization application

Cluster variations of the same document title and add a local document identifier.

You are assigning the following variation:

Variation ID:	50935
Variation Name:	TREATY OF COMMERCE AND NAVIGATION BETWEEN GREAT BRITAIN AND JAPAN Signed at London, 16th July, 1894 Ratifications Exchanged at Tokyo, 25th August, 1894
Origin image:	
Saved Cleaned name:	None
Page context:	Load whole page

The following unassigned variations are similar:

1.00	<input checked="" type="checkbox"/> TREATY OF COMMERCE AND NAVIGATION BETWEEN GREAT BRITAIN AND JAPAN Signed at London, 16th July, 1894 Ratifications Exchanged at Tokyo, 25th August, 1894	Page context	<input type="checkbox"/>	Uncertain	
0.99	<input checked="" type="checkbox"/> TREATY OF COMMERCE AND NAVIGATION BETWEEN GREAT BRITAIN AND JAPAN. Signed at London, 16th July, 1894. Ratifications Exchanged at Tokyo, 25th August, 1894	Page context	<input type="checkbox"/>	Uncertain	
0.99	<input checked="" type="checkbox"/> TREATY OF COMMERCE AND NAVIGATION BETWEEN GREAT BRITAIN AND JAPAN Signed at London, 16th July, 1894 Ratifications Exchanged at Tokyo, 25th August, 1894	Page context	<input type="checkbox"/>	Uncertain	
0.73	<input type="checkbox"/> TREATY OF COMMERCE AND NAVIGATION BETWEEN GREAT BRITAIN AND JAPAN Signed at	Page context	<input type="checkbox"/>	Uncertain	

852.349/pct:33/0/default.jpg ril, 1911

Additional classification and relations

Dictionary based prediction of classification of document type and parties involved.

Extraction of date of signature using **regex**.

Manual **validation** of predicted values.

Relations between documents added manually.


1	Legislation	Treaty	Supplement
3	Orders	TREATY	Additional
4	Act	between	PURSUANCE
5	Ordinance	CONVENTION	Annexed
6	Instructions	Convention	completing
7	Regulations	Articles Between	annexed
8	laws	TREATIES	Annexo
9	RULES	JOINTLY DETERMINED	ADDITIONAL
10	Fees	DECLARATION	Supplementary
11	Order	Agreement	SUPPLEMENTARY
12	Rules	AGREEMENT	AMENDED
13	SCALE OF COMMISSIONS	PROTOCOL	Annex
14	CODE	AGREEMENTS	Treaty Revision
15	ORDER	AGREEMENT	
16	IN COUNCIL	Treaty	
17	REGULATIONS		
18	ORDER		
19	Court Fees		


The result

Automatically generated Invenio based repository available at <https://adc-treaties.unibas.hasdai.org/> . It contains 2400+ instances of documents that are mapped to less than 400 canonical documents.

Filters:

- Parties
- Year published
- Year signed
- Document type

 Universität
Basel



parties

☐ UK (213)

☐ China (207)

☐ Japan (101)

☐ USA (39)

☐ Corea (36)

☐ France (32)

☒ Siam (31)

☐ Russia (20)

☐ Germany (17)

☐ Portugal (13)

published

☐ 1913 (107)

☐ 1930 (103)

☐ 1907 (102)

☐ 1923 (101)

☐ 1920 (100)

☐ 1927 (100)

☐ 1902 (97)

☐ 1903 (97)

☐ 1905 (94)

☐ 1901 (85)

signed

☐ 1858 (17)

☐ 1928 (16)

Found 31 results.

< 1 2 >

GENERAL REGULATIONS UNDER WHICH BRITISH TRADE IS TO BE CONDUCTED IN SIAM, IN CONFORMITY WITH THE TREATY CONCLUDED BETWEEN HER BRITANNIC MAJESTY AND THE KINGS OF SIAM

Parties: UK, Siam

Signed: 1860

Type: Legislation

TREATY PORTS, PORTS OF CALL, AND PLACES OPEN TO FOREIGN TRADE IN THE FAR EAST

Parties: China, Siam, Japan, Corea

Type: General Guidelines

TREATY OF AMITY, COMMERCE, AND NAVIGATION, BETWEEN THE GERMAN CONFEDERATION AND SIAM.

Parties: Siam, Germany

Signed: 1862

Type: Treaty

Tariff of Export and Inland Duties to be levied on Articles of Trade

Parties: UK, Siam

Signed: 1856

Type: Supplement

RULES AND REGULATIONS FOR THE PEACE, ORDER, AND GOOD GOVERNMENT OF HER MAJESTY'S SUBJECTS BEING WITHIN THE DOMINIONS OF THE KINGS OF SIAM.

Parties: UK, Siam

Type: Legislation

Protocol

Parties: Siam, Japan

Signed: 1898

Type: Treaty

LAWS CONCERNING VESSELS BELONGING TO SIAM, AND VESSELS FROM FOREIGN PORTS, LARGE VESSELS AND LIGHTERS, WHICH COME INTO THE CHOW PHYA RIVER, OR INTO ANY OF THE RIVERS OF THE PROVINCES BELONGING TO SIAM


Parties: UK, Siam

Summary


1. Large corpus (100k+ pages)
2. Extract the relevant pages (Treaties Sections)
3. Annotate headers
4. Process the annotations: Normalize, Categorize and relate
5. Generate Invenio repository

Key technologies: iiif (mirador), ORCID, invenio, Abbyy Finereader, Google Spreadsheets

Institutions: Basel University, Data Futures, Notre Dame University



Universität
Basel



parties

☐ UK (213)

☐ China (207)

☐ Japan (101)

☐ USA (39)

☐ Korea (36)

☐ France (32)

☒ Siam (31)

☐ Russia (20)

☐ Germany (17)

☐ Portugal (13)

published

☐ 1913 (107)

☐ 1930 (103)

☐ 1907 (102)

☐ 1923 (101)

☐ 1920 (100)

☐ 1927 (100)

☐ 1902 (97)

☐ 1903 (97)

☐ 1905 (94)

☐ 1901 (85)

signed

☐ 1858 (17)

☐ 1928 (16)

Found 31 results.

<

1

2

>

GENERAL REGULATIONS UNDER WHICH BRITISH TRADE IS TO BE CONDUCTED IN SIAM, IN CONFORMITY WITH THE TREATY CONCLUDED BETWEEN HER BRITANNIC MAJESTY AND THE KINGS OF SIAM

Parties: UK, Siam
Signed: 1860
Type: Legislation

TREATY PORTS, PORTS OF CALL, AND PLACES OPEN TO FOREIGN TRADE IN THE FAR EAST

Parties: China, Siam, Japan, Korea
Type: General Guidelines

TREATY OF AMITY, COMMERCE, AND NAVIGATION, BETWEEN THE GERMAN CONFEDERATION AND SIAM.

Parties: Siam, Germany
Signed: 1862
Type: Treaty

Tariff of Export and Inland Duties to be levied on Articles of Trade

Parties: UK, Siam
Signed: 1856
Type: Supplement

RULES AND REGULATIONS FOR THE PEACE, ORDER, AND GOOD GOVERNMENT OF HER MAJESTY'S SUBJECTS BEING WITHIN THE DOMINIONS OF THE KINGS OF SIAM.

Parties: UK, Siam
Type: Legislation

Protocol

Parties: Siam, Japan
Signed: 1898
Type: Treaty

LAWS CONCERNING VESSELS BELONGING TO SIAM, AND VESSELS FROM FOREIGN PORTS, LARGE VESSELS AND LIGHTERS, WHICH COME INTO THE CHOW PHYA RIVER, OR INTO ANY OF THE RIVERS OF THE PROVINCES BELONGING TO SIAM

Parties: UK, Siam