KanripoX: A tagset for connecting digital texts

Christian Wittern

February 14, 2021

1 Abstract

Our cultural heritage is based to a large degree on texts. Every one of these texts is part of a network of ideas, allusions, quotations, commentaries, criticisms, improvements, adaptions and many other ways of relating. Some of these relations can be described at the level of metadata, such as in library catalogs, or by classifying them in categories. In this paper, I want to describe a way to establish relations between texts on a much lower level, to align corresponding passages as a way to support the establishment of intertextual networks.

This work originated in the practical problem of combining two existing repositories of premodern Chinese texts and is an attempt to solve this in a way that can be generalized to other projects.

2 The Kanseki Repository

The original version of the Kanseki Repository was developed since 2013 by a research group at the Center for Informatics in East-Asian Studies, Institute for Research in Humanities, Kyoto University. The website serving as one of the interfaces to the Repository was opened to the public in March 2016. Since then minor revisions and developments have been carried out, and new texts have been added regularly¹.

2.1 Overall structure

The Kanseki Repository (KR) consists of three components:

- Text repositories at GitHub: @kanripo/*
- A user facing webserver accessed through www.kanripo.org
- An extension to Emacs Orgmode called Mandoku² (as of January 2021, there have been 1350 downloads of the module at https://melpa.org/#/mandoku since October 2014)

¹For more information on the Kanseki Repository, please see ウィッテルン・クリスティアン (編): センター研究年報2015 特集 漢籍リポジトリ / Special issue Kanseki Repository (ed : Christian Wittern) CIEAS Research Report 2015, available at http://hdl.handle.net/2433/210140, as well as http://blog.kanripo.org.

²More information about Mandoku is availabe in *Special issue Kanseki Repository* p. 44-55.

🐺 Kanseki Repository	/ Catalog Blog	Title Search	Text Search	Cont
漢リポ Kanseki Repository Electronic texts for researchers from researchers.				
News	About Kanseki Repository			
 i2016-03-10 First public release of the website. i2015-10-22 	This website provides access to the texts availa the moment, there is only very simple search fur the users.	ble in the GitHub user accoun nctionality, but it is planned to e	t @kanripo (Kanseki Reposito extend this depending on the ne	ry) . At eds of
New website available, more than 9000 texts	More information about the Kanseki Repository,	, its design considerations and	aims can be found here: Kanripo	b Blog.
online. • [2015-05-06] Started move to	A more sophisticated and powerful way to acce Emacs.	ess the repository is through Ma	Indoku, an extension of Org-Mo	de for
github. Some Daoist texts are up for testing	Catalog			
purposes.	KR1 經部			
First pretest release of	KR2 史部			
the Kanseki Repository.	KR3 子部			
	KR4 集部			
	KR5 道部			
	KR6 佛部			

Figure 1: The Kanseki Repository website

The texts in the KR are arranged in 6 main categories and 81 subcategories. The main categories are as follows:

- KR1 經部 Jing bu Confucian Classics (incl. music, dictionaries and elementary learning)
- KR2 史部 Shi bu Historiography and politics
- KR3 子部 Zi bu Masters, philosophers and treatises
- KR4 集部 Ji bu Anthologies (Poetry and Collected Writings)
- KR5 道部 Dao bu Daoist texts
- KR6 佛部 Fo bu Buddhist texts

For every text item, there might be multiple witnesses of the text, in addition, there is at least one version of the text curated by the editors of the KR. For many texts, in addition to a transcribed version of the text, there are also digital facsimiles of the source editions. Providing reliable editions of premodern Chinese texts in a coherent and simple format with a free license was the main reason for starting the project.

2.2 Website

For the benefit of readers not familiar with the Kanseki Repository web site, here is a quick overview with some screenshots.

The landing page of the website is shown in Figure 1, the main categories are available for browsing and search can be conducted for titles or in the texts.

In Figure 2, the first page of the results for a search is displayed, as can be seen, the results are displayed as KWIC (Keyword in Context) table, ordered by the characters

🤖 漢籍リポジト	リ目録題名検索	テキスト検索	連絡先 サ	インイン
「 <mark>七經</mark> 」を検索	した。 合計1685内1から20の表示。	テキスト	毎まとめ	て表示
Sort - テキスト番号順 - 日付順 - 検索語順 - 検索語の前の字で 朝代 - 唐 448 - 清 300 - 宋 290 - 明 191	KR6e0158 依楞嚴究竟事懺-明-禪修, 1, 519a KR6e0071 華嚴經行顏品疏鈔-唐-澄, 1, 204a KR3e0096 續名醫類案-清-魏之琇, 31, 17b KR5b0150 靈寶領教濟度金書寧全, 152, 3a KR2d0006 通志-宋-鄭樵, 123, 46a KR3k0004 北堂書鈔-唐-虞世南, 95, 1b KR3k0013 册府元龜-宋-王欽若, 563, 10b KR3k0055 御定淵鑑類図-清-張英, 165, 3b KR3k0055 御定淵鑑類図-清-張英, 197, 11a KR2d0011 古今紀要-宋-黃壽, 3, 12a	辭如第七疊"南無佛"南無法 結歸"七疊"四展卷難思*初 婦人七七經暫無子汝母禀厚之 風靜煥七經等景之功隂返陽回 也乃按七經識明堂圖河間古辟 也乃按七經識明堂圖河間古路 也乃按七經藏明堂圖河間古路 五純案七經藏明堂圖河間古路 五純案七經藏明堂剛河間古路 五純案七經藏明堂河間古路	(夾注) (夾注) (夾注)	
 ・ 佛部 637 ・ 史部 276 ・ 子部 259 ・ 道部 219 ・ 集部 202 ・ 観部 91 	KR2d0022後漢書補逸-清-姚之駰,19,8a KR2l0003 麟臺故事-宋-程俱,2,11a KR6n0007 瑜伽論記·唐-通倫,21,784b KR5g0030上清靈寶大法(一)→王,2,1a KR6s0095 大唐開元釋教廣品歷章-唐,20,439 KR6s0093 開元釋教錄-唐-智昇,1,480a KR6s0097 貞元新定釋教目錄-唐-圓,1,777b	 純傳有七輕識註云七輕詩書禮 語爾雅七輕號義凡一百六十五 訖至第七經一僧祗修無相行方 法而有七經一則洞視百靈吞服 4 行三十七經一卷*比丘避女惡 行三十七經一卷*犯戒報應輕 行三十七經一卷*犯戒報應輕 	(夾注) (異本)	
部/類 - 佛部/事彙部類 236 - 佛部/華嚴部類 133 史部/目錄類 111 - 子部/項書類 109 - 道部/太玄部 91 - 集部/模書類 77	KR6s0093 開元釋教錄。唐-智昇, 20, 694b KR6s0097 貞元新定釋教目錄-唐-圓, 30, 1042 KR6s0093 開元釋教錄-唐-智昇, 4, 518a	行三十七曜一卷三紙•比丘避 a 行三十七繩一卷三紙•比丘避 •七女七繩一卷右一十五部二	(異本) (異本) (異本)	

Figure 2: Results of the search for the term 七經 qijing

of the search string and the immediately following characters. Other sort orders and a subset of the search result based on publication time or classification within the repository can be selected from the menu on the left side.

Figure 3 finally shows a text, in this case both the transcribed version and the digital facsimile.

Users can log into to the website with their GitHub user id, this will give access to user customizations, for example in the display of search results. Users can also maintain text collections that allow them to limit search results to specific texts of interest. All user data are stored in repositories in the GitHub account of the user, none is held on the website. Users can also clone texts from the @kanripo account to their own account and edit the texts there. This is also the preferred route for reporting mistakes in the texts.

2.3 Interfaces

In addition to the website, the text of the KR can be accessed through the GitHub API. The KR website also offers an API, which provides access to the metadata and search interface.

2.4 User community

As of January 2021, there are 9351 texts available in the Kanseki Repository. These texts can be accessed through GitHub, either through the GitHub website, or through the Application Program Interface (API). The main access for users is through the website, a tiny group of users is also accessing the site through Mandoku.

• There are 195 registered users who logged into the Kanseki Repository website at least once.

1001-011 基 单 递 法 动 道 中 上 太 X 达卡猫爆燃专之中七、 无论太漏品第一 基 单 递 法 动 道 中 上 太 MW 基 单 递 法 动 道 中 上 太 MW 基 单 递 法 动 道 中 上 太 MW 基 单 递 法 动 道 中 上 太 MW 基 单 递 法 动 道 中 上 太 MW 基 单 递 法 动 道 中 上 太 MW 基 单 递 法 动 道 中 上 太 MW 基 年 递 法 动 道 中 上 太 MW L 和 Systeman L 和 L 中 D 和 Change L 上 MW D 和 Systeman D 和 DO1-001 D 和 D 和 Systeman L 和 D 和 Systeman D 和 D 和 D D → Systeman		X +
Liunia de la certa de	[001-001a]	
元始大羅品第一 WH 元始大羅品第一 WH 元始大羅品第二 WH 和時 元始大羅品第二 WH 和時 五人名日左玄長跪道前上白 其名石白左玄長跪道前上白 天尊帕名十極五無間罪是時會中有一 真人名日左玄長跪道前上白 天尊有名日在一 門滿滿說名日石一十極一者。太上 門海濤說名日日一十極一者。大東 中道妙 光一一 「一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一	太上中道妙法蓮華經卷之一杜一	經華蓮法妙道中上太
WF 元始天尊音於大羅天上廣化天宫會諸天 仙樂說 1 極照五無間罪是時會中有一 友人名日左玄長驗道前上白 天尊向名十極五無間罪於是 天尊的名相 無極滿空界樂生未曾聞我會諸仙樂廣演 解脫何會散地就告悔代、露罷從空 [001-001b] 下日月躍塗合間仙解脫語真仙大喜悅即 便隨仙人供給於所須探薪及藥臟隨時供 給與情存妙法衆戰心無懈倦音爲諸樂生 動修於大法願保罵己身及以五谷藥故爲 大國王所求聽此法遂致得成仙令故爲汝 該爾時 天尊告諸仙樂吾今爲汝說此妙法無量樂 生皆得解脫即時諸天仙樂悉以散花末味 上妙寶香旃懶沉水則真衣服要強儵觸音 蓋伎樂空歌頌礼供養 [001-002] 七寶瑶臺十方仙聖無量樂生皈依仙道是 時 元大寶西說 四一一一	元始大羅品第一	
元始天萼昔於大羅天上廣化天宫會諸天 仙樂說十極解脱五無間罪是時會中有一 真人名曰左玄長幾道前上白 天萼內留十種五無間罪於是 天尊於偈繡曰 無極滿空界樂生未曾聞我會諸仙樂廣演 解脫門停酸地獄苦今悟天仙人歌臨從空 [001-001b] 下日月癰瓊台聞仙解脫語真仙大喜悅即 價隨仙人供給於所須採荔及葉藏脑時供 紛與情存妙法衆勤心無懈倦普萬脑樂生 勤修於大法願保罵已身及以五谷藥故爲 大國王所求護此法遂致得成仙今故爲汝 腔釋時 天尊告諸仙樂吾今爲汝說此妙法無量衆 生皆得解脫即時諸天仙樂悉以散花末味 上妙實香游橇沉水野具衣服惡強幢觸 若文響時閒 天寶告諸仙樂吾今爲汝說此妙法無量衆 生皆得解脫即時諸天仙樂悉以散花末味 上妙實香游橇沉水野具衣服惡強幢觸 若文響時閒 天寶告諸仙樂吾今爲汝說此妙法無量衆 生皆得解脫即時諸天仙樂悉自 麗子方仙聖無量樂生皈依仙道是 時 元太寶兩寇 百	爾時	解血玉天白 仙云阙 大
山梁航十極解脱五無間罪是時會中有一 真人名日左玄長跪道前上白 天萼初名十楹照間無影響是 天萼就偈離日 無極減至界衆生未曾聞我會諸仙梁廣演 解脱門停酸地獄苦今悟天仙人歌謡從空 [01-001b] 下日月糶瓊台間仙解脫語真仙大喜悅即 僅随仙人供給於所須採薪及葉蘿随時供 給奧情存妙法衆難心無懈倦嘗爲諸索生 對修於大試靈際係高三身及以五谷藥故爲 大國王所求護此法遂致得成仙令故爲汝 訖獨時 天萼告諸仙衆吾今爲汝說此妙法無量衆 生皆得解脫即時諸天仙衆悉以散花木昧 上妙寶香旃欄沉水卧具衣服瓔珞幢幡寶 盖伎樂空歌頌礼供悉 [01-002] 下日月糶瓊台間如解脫語真仙大喜悅即 伊蘭仙人供給於所須採薪及葉蘿随時供 對修於大誠爾係高三身及以五谷藥故爲 大國王所求護此法遂致得成仙令故爲汝 訖爾時 天尊告諸仙衆吾今爲汝說此妙法無量衆 生皆得解脫即時諸天仙衆悉以散花木昧 上妙寶香旃欄沉水卧具衣服瓔珞幢幡寶 盖伎樂空歌頌礼供悉 [01-002] [元始天尊昔扵大羅天上廣化天宫會諸天	間部金金丹 個儿間 八
真人名日左玄長跪道前上白 門滿說何名。說天。元中道 天萼的偈蕭日 門滿說何名。說天。元中道 無極滿空界衆生未曾聞我會諸仙衆廣演 門滿說何名。說天。元中道 解脫門停酸地獄苦今悟天仙人歌謡從空 吃家。當十本些十大城。 [001-001b] 一 原随仙人供給於所須採薪及奠蘿隨時供 一 給奧情存妙法衆勤心無懈倦曾爲諸衆生 一 鄭修於大法願保高己身及以五谷藥故爲 一 大國王所求護此法遂致得成仙今故爲次 一 天萼古諸仙衆吾今爲次說此妙法無量祭 一 生皆得解脫即時諸天仙衆悉以敗花末味 一 上妙寶齋筋爛沉水皆具衣服璎珞儵幗寶 一 蓋伐樂空歌颂礼供養 一 [01-002a] 一 [01-002a] 一 [01-002a] 一 古安市 方大萼商前人 一 [01-02a] 一 古安市 一 「方」一 一 「方」一 一 「方」一 一 「白」 一 「白」 一 「方」 一 「方」 一 「方」 一 「一 一 「方」 一 「方」 一 「方」 一 「古」 一 「方」 一	仙衆說十極解脫五無間罪是時會中有一	脱極导导人 化始时 上
天萼何名十極五無問罪於是 「停空宿名白 十葉 结道 英華統陽空界衆生未曾聞我會諸仙衆廣演 停空宿者十左士 極 昔 大妙 無極滿空界衆生未曾聞我會諸仙衆廣演 「停空宿者十左世 種 古 大妙 解版門停酸地獄苦今悟天仙人歌謡從空 地 泉 口 極 玄 一解代 然 羅達 [001-0016] 一 下日月攤瓊台聞仙解脱語真仙大喜悅即 一 伊蘭仙人供給於所須採薪及葉蘆隨時供 苦。一般 完 長 脫大 品 靠 給與情存妙法衆勤心無懈倦普爲諸衆生 一 勤修大法願保局已身及以五谷藥故爲 一 大國王亦或進法遂致得成仙今故爲汝 一 大國王亦或進法遂致得成仙今故爲汝 一 大國王亦或進法遂致得成仙今故爲汝 一 大方音音 一 大大 二 大大 二 大大 二 「一 一 「日 二 「日 二 「一 一 「一 二	真人名曰左玄長跪道前上白	門滿該何名 說天 元中
天尊說偈諸日 行空宿石口丁草、加口丁草、加口丁草、加口丁草、加口丁草、加口丁草、加口丁草、加口丁草、加	天尊何名十極五無間罪扵是	适应值夕日 上貧 払送
無極滿空界衆生未曾聞我會諸仙衆廣演 解脫門停酸地獄苦今悟天仙人歌謡從空 1001-001b 一下日月耀瓊台聞仙解脫語真仙大喜悅即 便隨仙人供給於所須採薪及葉藏隨時供 給與情存妙法衆勤心無懈倦普爲諸衆生 勤修於大法顯保爲已身及以五谷藥故爲 大國王所求護此法遂致得成仙今故爲汝 說爾時 天萼告諸仙衆吾今爲汝說此妙法無量衆 生皆得解脫即時諸天仙衆悉以散花未味 上妙寶香旆嵲沉水卧具衣服瓔珞幢幡寶 盖伎樂空歌頌礼供養 [01-002a] 一 七寶瑶臺十方仙聖無量衆生皈依仙道是 時 元治天尊而說偈言	天尊說偈譛曰	行至個石口 丁素 石退
 解脫門停酸地獄苦今悟天仙人歌謡從空 [001-001b] 下日月耀瓊台聞仙解脫語真仙大喜悅即 使隨仙人供給於所須採薪及葉藏隨時供 給與情存妙法衆勤心無懈倦普爲諸衆生 勤修於大法願保爲已身及以五谷藥故爲 大國王所求護此法遂致得成仙今故爲汝 大國王所求護此法遂致得成仙今故爲汝 大國王所求護此法遂致得成仙今故爲汝 大國王所求護此法遂致得成仙今故爲汝 大國王所求護此法遂致得成仙今故爲汝 大國王所求護此法遂致得成仙會故爲汝 大國王所求護此法遂致得成仙會故爲汝 大國王所求護此法遂致得成仙會故爲汝 大國王子 做任王思 在無見也 中書書 大四王子 做任王思 在無見也 	無極滿空界衆生未曾聞我會諸仙衆廣演	酸吊潜十左柱极苷 大妙
[001-001b]下日月耀瓊台間仙解脫語真仙大喜悅即 便隨仙人供給於所須探薪及葉蓏隨時供 給與情存妙法衆勤心無懈倦曾爲諸衆生 勤修於大法顯保爲已身及以五谷藥故爲 大國王所求護此法遂致得成仙今故爲汝 訖爾時天萼告諸仙衆吾今爲汝說此妙法無量衆 生皆得解脫即時諸天仙衆悉以散花末味 上妙寶香旃標沉水卧具衣服瓔珞幢幡寶 蓋伎樂空歌頌礼供養乙的1-002al 时 无方聲而訪協信元为天尊而訪協信 元方葉南市訪協信	解脫門停酸地獄苦今悟天仙人歌謡從空	执货日极方 解决 羅注
FEJJ羅瓊台間仙解脫語真仙大喜悅即 減生 五夜 脱大 中連 便隨仙人供給於所須採薪及葉蘆隨時供 給與情存妙法衆勤心無懈倦普爲諸衆生 添生 五夜 脱大 中連 勤修於大法願保爲已身及以五谷藥故爲 天離先 五麗 第 第 大國王所求護此法遂致得成仙今故爲汝 市 前 前 十 差 太國王所求護此法遂致得成仙今故爲汝 市 一 一 差 太國王所求護此法遂致得成仙今故爲汝 市 元 一 差 太國王所求護此法遂致得成仙今故爲汝 市 元 一 差 大國王所求護此法遂致得成仙今故爲汝 市 元 一 差 大國王所求護此法遂致得成仙今故爲汝 市 元 一 一 大國王所求護此法遂致得成仙今故爲汝 市 一 一 一 大國王常常 前標示水 上 第 一 一 大國王常常 小 ① 一 一 大國王常常 市 一 一 一 大國王常常 市 一 一 一 大國王常常 市 一 一 一 大國王常常 市 市 社 一 大國王常常 近 一 一	[001-001b]	
@ 随仙人供給於所須探薪及集蓏随時供 ぎ、菜、魚、菜、白、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、、	下日月耀瓊台聞仙解脫語真仙大喜悅即	颜生 五衣 脫太 即連
 給與情存妙法衆勤心無懈倦普爲諸衆生 勤修於大法願保爲已身及以五谷藥故爲 大國王所求護此法遂致得成仙今故爲汝 試爾時 天萼告諸仙衆吾今爲汝說此妙法無量衆 生皆得解脫即時諸天仙衆悉以散花末味 上妙寶香旃標沉水卧具衣服瓔珞幢幡寶 盖伎樂空歌頌礼供養 [001-002a] [001-002a] [001-002a] [1]道 無天 一經 一時天 心會 是白 是化 一 一時天 心 一時天 戶 一時天 前諸 杜 空演 一天 一 	便隨仙人供給扵所須採薪及菓蓏隨時供	苦禾 無跪 五籬 第菙
勤修於大法願保爲已身及以五谷藥故爲 大國王所求護此法遂致得成仙今故爲汝 訪爾時 天尊告諸仙樂吾今爲汝說此妙法無量衆 生皆得解脫即時諸天仙樂悉以散花末味 上妙寶香旃標沉水卧具衣服璎珞幢幡寶 盖伎樂空歌頌礼供養 [001-002a] 時 七寶瑶臺十方仙聖無量衆生皈依仙道是 時 元始天尊而說偈言 世界王之,從在王思,在無耳也	給與情存妙法衆勤心無懈倦普爲諸衆生	尽自 崩資 無王 一颏
大國王所求護此法遂致得成仙今故爲汝 「「「」」」」」」」」 「「」」」」」」 「「」」」」」 「「」」」」」 「」」」」 「」」」」 「」」」」 「」」」」 「」」」」 「」」」」 「」」」」 「」」」」 「」」」」 「」」」」 「」」」」 「」」」」 「」」」」 「」」」」 「」」」」 「」」」」 「」」」」 「」」」」 「」」 「」」」 「」」 「」」 「」」 「」」 「」」 「」」 「」」 「」」 「」」 「」」	勤修扵大法願保爲己身及以五谷藥故爲	
試爾時 天尊告諸仙樂吾今爲汝說此妙法無量衆 生皆得解脫即時諸天仙樂悉以散花末味 上妙寶香旃槇沉水卧具衣服璎珞幢幡寶 益伎樂空歌頌礼供養 [001-002a] ■ 七寶瑶臺十方仙聖無量衆生皈依仙道是 時 元始天尊而說偈言	大國王所求護此法遂致得成仙今故爲汝	作剧 悲剧 图上 卷
天萼告諸仙衆吾今爲汝說此妙法無量衆 仙會 是白 是化 一 生皆得解脫即時諸天仙衆悉以散花末味 人諸 一時天 一 上妙寶香旃檀沉水卧具衣服璎珞幢幡寶 盖伎樂空歌頌礼供養 〇1-002a) 一 ● ● ● ● ● ● ● ● ● ● ● ● ● ○ □<	說爾時	天我 农上 罪盾 之
生皆得解脫即時諸天仙衆恐以散花末味 上妙寶香旃檬沉水卧具衣服璎珞幢幡寶 盖伎樂空歌頌礼供養 [001-002a] 日 七寶瑶臺十方仙聖無量衆生皈依仙道是 時 元始天尊而說偈言 世界王主。姓代王思,在無耳也	天尊告諸仙衆吾今爲汝說此妙法無量衆	仙食 具白 星化 一
Lbg音游標沉水卧具衣服璎珞幢帽寶 盖伎樂空歌頌礼供養 [001-002a] 七寶瑶臺十方仙聖無量衆生皈依仙道是 時 元始天尊而說偈言 世界工工,始仕工界、在無耳中	生皆得解脫即時諸天仙衆悉以散花末味	人世 天日 天化
盖伎樂空歌頌礼供養 歌仙。 會宮 [001-002a] : : : : : 中會 七寶瑶臺十方仙聖無量衆生皈依仙道是 : : : : 中會 花房 要面的 : : : : : : 市 : : : : : : 大寶珉臺十方仙聖無量衆生皈依仙道是 : : : : : : 時 : : : : : : : 元始天尊而說偈言 : : : : : :	上妙寶香旃檩沉水卧具衣服瓔珞幢幡寶	公前 一切太 的
1001-002al <	盖伎樂空歌頌礼供養	歌仙 會宮
七寶瑶臺十方仙聖無量衆生皈依仙道是 花 丁 省 時 花焼天萼而說偈言 空演 一天	[001-002a]	经废 ③ 品瓜
	七寶瑶臺十方仙聖無量衆生皈依仙道是	
	時	化原 有諸 社
	元始天尊而說偈言	空演 一天 一
我高大主 催吐大介 任無局中	我爲天主 惟住天界 在無爲中	
尚爲法來 諸人云何 不勤爲法 HFL 001-001a	尚爲法來諸人云何不勤爲法	HFL 001-001a

Figure 3: Transcribed text and digital facsimile of KR5h0001 太上中道妙法蓮花經

- 39 Pull requests received until 2021-01-11
- Average count of monthly page accesses to the Kanseki Repository: 2 178 647

3 Thesaurus linguae sericae

The Thesaurus linguae sericae (TLS, Chinese name 漢學文典) is an electronic resource that combines premodern Chinese texts with translations and a lexical database with deep linguistic description and a unique grouping of words into an organized hierarchy of concepts³.

It has a long history, originating in a personal Filemaker database inspired by Christoph Harbsmeier and maintained by Jens Østergaard Petersen. Since 2015, efforts have been underway to provide a new interface and link it with the Kanseki Repository. The latter goal proved evasive so far, and this integration is the very topic of ongoing research, including this paper. So far, the included texts have been updated and expanded and part of the functionality is available now through the web interface at https://hxwd.org.

Key features of the TLS are 4:

³For more information about the TLS see Christoph Harbsmeier, "Philological Reflections on Chinese Conceptual History: Introducing Thesaurus Linguae Sericae", in: *Keywords in Chinese Culture*, ed. by Wai-yee Li and Yuri Pines, The Chinese University of Hong Kong Press, 2020.

⁴The following is partly based on: https://web.archive.org/web/20070609105645/http://www. sino.uni-heidelberg.de/database/tls/index.htm and http://ancientworldonline.blogspot. com/2009/11/open-access-historical-and-comparative.html for more information.

- Integration of source text with translation and analytical tools.
- With its focus on distinctive semantic nuances, it serves as a synonym dictionary of classical Chinese.
- It systematically organises the Chinese vocabulary in taxonomic and mereonomic hierarchies, thus showing up whole conceptual schemes or cognitive systems;
- It systematically registers a range of lexical relations like antonym, converse, epithet etc.; TLS thus aims to define the conceptual space in relation to other words.
- It incorporates detailed syntactic analysis of (over 1200 distinct kinds of) syntactic usage; TLS thus enables users to make a systematic study of such basic phenomena as the natural history of abstract nouns in China.
- TLS explores the conceptual schemes of premodern Chinese on the basis of a corpus of translated texts interlinked with an analytic dictionary.
- Text and dictionary are constantly held up against one another. Our understanding of the texts and the Chinese system of meanings can be refined by through this close confrontation.
- TLS associates Chinese concepts with concepts from the European antiquity, aiming to make the classical Chinese evidence comparable to that of other cultures.
- TLS seeks to make more precise the criteria used in translation classical Chinese, through a detailed description of the semantic relations that obtain among Chinese words.

3.1 TLS in action

In order to better understand the above, here is a series of screenshots that demonstrate some of these features.

In Figure 4 the beginning of the Lunyu is shown, with one of the translations available in the TLS displayed next to the source texts. All texts in the TLS are displayed with a phrase of the Chinese texts as the basic unit, translations are aligned to it. There are two columns to display translations, commentaries or other information related to the text passage in question. Registered users with the necessary permissions can also suggest corrections to the translations or translate new passages.

Figure 5 shows the same passage with the annotations brought in view. Annotations follow the line they comment on, with details of the lemma, the concept this is assigned to, the syntactical function and a translation gloss for the lemma. The buttons to the right of the annotations are used for peer review and comments.

The TLS allows its users to inspect the concepts a given character has already been assigned to (such a combination of character, concept and syntactical function is called a syntactic word, SW in the TLS), this gives an overview of the range of registered semantic fields. In Figure 6 this view is actived for the character $\stackrel{\text{$\square$}}{=}$, which is done by selecting the character. The list of concepts is shown to the right sight. Registered users also use this screen to add new annotations. The numbers to the right show the number of assigned syntactic words within a given concept, for $\stackrel{\text{$\square$}}{=}/STUDY$ for example, there are

海	冬女女 TLS Browse ▼ 1	論語 目録 ▼ 💿 Source: CHANT Bookmarks ▼ 內部 ▼	Search
	1 《學而篇第一》 line 1 / 0%	Translation by C. Harbsmeier 🕶	Translation by TLS Project -
	1 《學而篇第一》		
	 1.1子曰: 「學而時習之, 不亦說乎? 有朋自遠方來, 不亦樂乎? 人不知而不慍, 不亦君子乎?」 1.2有子曰: 「其為人也孝弟,而好犯上 	The Master said, "Having studied something to exercise it in practice, at the proper times, isn't that satisfying? To have a colleague come from a distant region isn't that delightful? When others do not appreciate one not to feel offended isn't that behaving in a gentlemanly fashion?" Yǒuzi said, "That anyone whose personality is both filial and fraternal and yet should	
	者,	be prone to offend the authorities,	
	^{新天,,} 不好犯上,而好作亂者,	For someone not prone to offending superiors yet to be prone to creating political rebellions,	
	未之有也。	that never occurs.	
	君子務本,	The gentleman concentrates his efforts on what is basic.	
	本立而道生。	Once what is basic is established, then the Way will [naturally] emerge.	
	孝弟也者,	As for filial piety and fraternal love (in general),	
	其為人之本與!」	these must count as the basis for Goodness, mustn't they?"	

Figure 4: First chapter of the Lunyu with translation

1 《學而篇 line 1 / 0%	第一》	Translation by C. Harbsmeier 🕶	Translation by TLS Project -	
1 《學而	篇第一》			
1.1子曰:		The Master said,		
	子 (zǐ)	TEACHER	n[post-N] our master; person in authority [Note that the [post-N] can n in fact be made explicit without breaking the rules of the grammar! CH]	ever
「學而時習	之,	"Having studied something to exercise it	in practice, at the proper times,	
Rhet:	TRICOLON-2+1+C	RESCENDO		
	學 (xué)	STUDY	vt[oN] act devote oneself to study; be devoted to study; engage in	
	而 (ér)	AND	intellectual work; work to improve oneself morally/intellectually padV1.postV2 sequence e.g. 拜而受之 "bow and accept it" and then, thereupon; (often coordinates two verbal predicates with the same	ംര× and
	時 (shí)	TIMELY	subject) nadV appropriate at the appropriate time; when the need arises Note	∂@X 非
			時	∆0 ∆
	時 (shí)	OFTEN	nadV from time to time> sometimes; often; periodically; several times	; at
			regular intervals; all the time	¢0X
	之 (zhī)	PRONOUN	npro.postVt nonreferential an object, things; a place; something [cogn object; indefinite dummy object pronouns without reference: things, people; someone 以約失之者"Those who, through keeping to the mai thing, get things wrong"]	nate in 🏠 @X
不亦說乎?		isn't that satisfying?		
Rhet:	悅 / 說 (yuè / yuè) ISOCOLON+PARA	DELIGHT LLEL+REPETITIO-	vi subj=nonhuman be delightful	்0X



1 《學而篇第一》	Translation by C. Harbsmeier	Existing SW for 學	\otimes
line 1 / 0%	Translation by St Harboniolo	At: KR1h0004_tls_001-1a.4	
1 《學而篇第一》		Line: 「學而時習之,	
1.1子曰:	The Master said,	Use one of the following syntactic words (SW), create a	New SW ,add
「學而時習之,	"Having studied something to exercise it in practice, at the proper times,	an existing Concept to the word or create a New Conce	pt .
不亦說乎?	isn't that satisfying?		
有朋自遠方來,	To have a colleague come from a distant region	学 (XUE) COMMAND New SW SYN 1	
不亦樂乎?	isn't that delightful?	學 (XUÉ) DOCTRINE New SW SYN 1	
人不知而不慍,	When others do not appreciate one not to feel offended		
不亦君子乎?」	isn't that behaving in a gentlemanly fashion?"	學 (xué) EMULATE New SW 1	
1.2有子曰:	Yǒuzi said,		
「其為人也孝弟,而好犯上	"That anyone whose personality is both filial and fraternal and yet should		
者,	be prone to offend the authorities,	學 (xué) KNOW New SW SYN 1	
鮮矣;	that is quite rare.		
不好犯上,而好作亂者,	For someone not prone to offending superiors yet to be prone to creating	學 (xué) LEARN New SW SYN 4	
	political rebellions,		
未之有也。	that never occurs.	Y (xue) RESEMBLE New Word	
君子務本,	The gentleman concentrates his efforts on what is basic.	學 (xué) SCHOLAR New SW 1	
本立而道生。	Once what is basic is established, then the Way will [naturally] emerge.		
孝弟也者,	As for filial piety and fraternal love (in general),	學 (Xué) SCHOOL New SW SYN 2	
其為人之本與!」	these must count as the basis for Goodness, mustn't they?"		
1.3子曰:	The Master said,	4 (Xue) STODENT New Word	
「巧言令色,	"Those who make their speech crafty and their appearance artificially	學 (xué) STUDY New SW SYN 20	
	distinguished		
鮮矣仁!」	are rarely indeed Good."	學 (XUÉ) WISE New Word	

Figure 6: Concepts related to 學 xué

1 《學而篇第一》 line 1 / 0% 1 《學而篇第一》	Translation by C. Harbsmeier -	學 (xué) SCHOLAR New SW 1 學 (xué) SCHOOL New SW SYN 2
11子曰:	The Master said	學 (xué) STUDENT New Word
(1) (1) (1) (1) (1) (1) (1) (1)	"Having studied something to exercise it in practice, at the proper times, isn't that satisfying? To have a colleague come from a distant region isn't that delightful? When others do not appreciate one not to feel offended	學 (xué) STUDY New SW SYN 20 1. The dominant word is xué 學 (ant. jiào 教 "train teach")which refers primarily to studying or training under another person, and secondarily to the learning by heart texts. Very often, the word retains a tinge of immitation.
小亦名于于:」 1.2有子曰: 「其為人也孝弟,而好犯上 老	Your and benaving in a genuemany rashor? Your said, "That anyone whose personality is both filial and fraternal and yet should	2. Shī 飾 and cóng 從 refer to deciding to study under someone and treating him as one's teacher.
1979 (1975) (1975)	that is quite rare.	3. Kǎo 考 refers to an investigation of a subject in a "scientific" spirit.
不好犯上,而好作亂者,	For someone not prone to offending superiors yet to be prone to creating political rebellions,	4. Jiū 究 and yán 研 refers to an in-depth study of a subject, typically involving a fair amount of reflection.
未之有也。 カス改士	that never occurs.	
右于····································	Once what is basic is established, then the Way will [naturally] emerge.	5. Zhì 治 and gōng 攻 refer to the specialised study, typically of a certain text.
孝弟也者, 其為人之本與!」 1 3 3 日:	As for filial piety and fraternal love (in general), these must count as the basis for Goodness, mustn't they?"	6. Xí 習 refers to rehearsal of what one has learnt, through repetition of text and/or through enactment in practice
「巧言令色,	"Those who make their speech crafty and their appearance artificially distinguished	学 (xué) WISE New Word
鮮矣仁!」	are rarely indeed Good."	

Figure 7: Contrast of words within the concept STUDY

1 《學而篇第一》 1.1子曰: The Master said, 「圖而時習之, "Having studied something to exercise it in practice, at the proper times, 亦亦更? isn't that satisfying? 有期自遠方來, To have a colleague come from a distant region 亦亦果?? isn't that delightful? 人不知而不僵, When others do not appreciate one not to feel offended 亦亦君子?? isn't that behaving in a gentlemanly fashion?" 1.2有子曰: Youzi said, 「其為人也孝弟, 而好犯上 "That anyone whose personality is both filial and fraternal and yet should af, be prone to offend the authorities, wit is quite rare. 不好犯上, 而好作亂者, For someone not prone to offending superiors yet to be prone to creating political rebellions, 未之有也。 that never occurs. 君子務本, The gentleman concentrates his efforts on what is basic. 本立而道生。 Once what is basis for Goodness, mustn't they?" 1.3子曰: The Master said, 「巧言令色, "The Master said, 「巧言令色, The Master said, 「巧言令色, The Master said, 「方言令合, The Master said, 「方言令合, The Master said, 「方言令合, The Master said, 「方言令合,	1 《學而篇第一》 line 1/0%	Translation by C. Harbsmeier +	學 (xué) STUDY New SW SYN 20 nab act: the attempt to learn about things (typically from a
小水説子?isn't that satisfying?有朋自遠方來, 不亦樂子?To have a colleague come from a distant region不亦樂子?isn't that delightful?人不知而不愠, 不亦君子乎?」When others do not appreciate one not to feel offended不亦君子乎?」isn't that behaving in a gentlemanly fashion?"1.2有子曰: 其為人也孝弟,而好犯上Yôuzi said,「其為人也孝弟,而好犯上"That anyone whose personality is both filial and fraternal and yet should者, 	1 《學而篇第一》 1.1子曰: 「 <mark>學</mark> 而時習之,	The Master said, "Having studied something to exercise it in practice, at the proper times,	teacher) study; the pursuit of intellectual/moral self-development; learning Save Use SwL: 41 • nab.post-N : the study of the subject N Save Use SwL: 1
人不知而不愠,When others do not appreciate one not to feel offended• vadN : learned, dedicated to the pursuit of learning 博學之生不亦君子乎?」isn't that behaving in a gentlemanly fashion?"Yôuzi said,• vadN : learned, dedicated to the pursuit of learning 博學之生1.267日:Yôuzi said,"That anyone whose personality is both filial and fraternal and yet should• vi : be a person of education, be a person who has engaged in有人也多弟,而好犯上"That anyone whose personality is both filial and fraternal and yet should• vi : be a person of education, be a person who has engaged in有人力不知而不能表,be prone to offend the authorities,wit is quite rare.不好犯上,而好作亂者,For someone not prone to offending superiors yet to be prone to creating political rebellions,wit is basic.未之有也。that never occurs.wit (NU) conative: try to learn to perform a contextually determinate action Save Use SwL:1者子務本,The gentleman concentrates his efforts on what is basic.wit is basic is established, then the Way will [naturally] emerge.孝弟也者,As for filial piety and fraternal love (in general),wit(N) perfective: study successfully the contextually determinate skill N Save Use SwL:11.3子日:The kaster said,wit(N) perfective: study successfully the contextually determinate skill N Save Use SwL:11.3子日:The kaster said,wit(N) is study with; become a student of a contextually determinate person Save Use SwL:2	不小說乎? 有朋自遠方來, 不亦樂乎?	isn't that satisfying? To have a colleague come from a distant region isn't that delightful?	nab.post-N : the study of the studying person N Save Use SwL: 1
1.2有子曰: Yǒuzi said, 「其為人也孝弟,而好犯上 "That anyone whose personality is both filial and fraternal and yet should 者, be prone to offend the authorities, 鲜矣; that is quite rare. 不好犯上,而好作亂者, For someone not prone to offending superiors yet to be prone to creating political rebellions, 未之有也。 that never occurs. 君子務本, The gentleman concentrates his efforts on what is basic. 本立而道生。 Once what is basic is established, then the Way will [naturally] emerge. 孝弟也者, As for filial piety and fraternal love (in general), 其為人之本與!」 these must count as the basis for Goodness, mustn't they?" 1.3子曰: The who make their speech crafty and their appearance artificially distinguished 「巧言令色, "Those who make their speech crafty and their appearance artificially determinate person Save Use SwL: 2	人不知而不慍, 不亦君子乎?」	When others do not appreciate one not to feel offended isn't that behaving in a gentlemanly fashion?"	• vadN : learned, dedicated to the pursuit of learning 博學之士
者,be prone to offend the authorities,proper study; be a person who has studied properly Save Use鲜矣;that is quite rare.SWL: 2不好犯上,而好作亂者,For someone not prone to offending superiors yet to be prone to creating political rebellions,• vt(+V[0]) conative: try to learn to perform a contextually未之有也。that never occurs.君子務本,The gentleman concentrates his efforts on what is basic.• vt(oN) conative: devote oneself to learning N; devote oneself to本立而道生。Once what is basic is established, then the Way will [naturally] emerge.• vt(oN) conative: devote oneself to learning N; devote oneself to孝弟也者,As for filial piety and fraternal love (in general),• vt(oN) perfective: study successfully the contextually其為人之本與! Jthese must count as the basis for Goodness, mustn't they?"• vt(oN) perfective: study successfully the contextually1.3子目:The Master said,• vt(oN) : study with; become a student of a contextually「巧言令色,"Those who make their speech crafty and their appearance artificially distinguished• vt(oN) : study with; become a student of a contextually	1.2有子曰: 「其為人也孝弟,而好犯上	Yǒuzi said, "That anyone whose personality is both filial and fraternal and yet should	• vi : be a person of education, be a person who has engaged in
 不好犯上,而好作亂者, For someone not prone to offending superiors yet to be prone to creating political rebellions, 未之有也。 that never occurs. 本立而道生。 Once what is basic is established, then the Way will [naturally] emerge. 孝弟也者, As for filial piety and fraternal love (in general), 其為人之本與!」 these must count as the basis for Goodness, mustn't they?" 1.3子曰: The Master said, 「巧言令色, "Those who make their speech crafty and their appearance artificially distinguished vt(oN) : study with; become a student of a contextually determinate person Save Use SwL: 1 	者, 鮮矣;	be prone to offend the authorities, that is quite rare.	proper study; be a person who has studied properly Save Use SWL: 2
末之有也。 that never occurs. 君子務本, The gentleman concentrates his efforts on what is basic. 本立而道生。 Once what is basic is established, then the Way will [naturally] emerge. 孝弟也者, As for filial piety and fraternal love (in general), 其為人之本與!」 these must count as the basis for Goodness, mustn't they?" 1.3子曰: The Master said, 「巧言令色, "Those who make their speech crafty and their appearance artificially distinguished determinate person Save Use SwL: 2	不好犯上,而好作亂者,	For someone not prone to offending superiors yet to be prone to creating political rebellions,	 vt(+V[0]) conative: try to learn to perform a contextually determinate action Save Use SwL: 1
本立而道生。 Once what is basic is established, then the Way will [naturally] emerge. the practice of N Save Use SwL:1 孝弟也者, As for filial piety and fraternal love (in general), 其為人之本與!」 these must count as the basis for Goodness, musth't they?" 1.3子曰: The Master said, "Those who make their speech crafty and their appearance artificially distinguished used use SwL:1 · vt(oN) perfective: study successfully the contextually determinate skill N Save Use SwL:1 · vt(oN) is study with; become a student of a contextually determinate person Save Use SwL:2	未之有也。 君子務本,	that never occurs. The gentleman concentrates his efforts on what is basic.	• vt(oN) conative: devote oneself to learning N; devote oneself to
其高人之本與 ! 」 these must count as the basis for Goodness, must in they? 1.3子曰: The Master said, 「巧言令色, "Those who make their speech crafty and their appearance artificially distinguished determinate skill N [save] Use [swL:1] ・ vt(oN) : study with; become a student of a contextually determinate person [Save] Use [swL:2]	本立而道生。 孝弟也者, 其改,人立主席,	Once what is basic is established, then the Way will [naturally] emerge. As for filial piety and fraternal love (in general),	the practice of N Save Use SWL: 1 vt(oN) perfective: study successfully the contextually
distinguished determinate person save Use swL: 2	具為人之本與!」 1.3子曰: 「巧言会色	these must count as the basis for Goodness, mustrit they?" The Master said, "Those who make their speech crafty and their appearance artificially	determinate skill N save Use swL: 1 • vt(oN) : study with: become a student of a contextually
鮮矣仁!」 are rarely indeed Good."	鮮矣仁!」	distinguished are rarely indeed Good."	determinate person save Use swL: 2

Figure 8: Syntactic words defined for STUDY/學 xué

already 20 SW registered. Clicking on these numbers will reveal the list of SW as shown in Figure 8.

The button in turquoise labelled "SYN" to the left of these numbers is contrasting the different characters that are used to cover various aspects of the concept, these contrasts for the concept STUDY are shown in Figure 7.

The list of SW in Figure 8 also gives the number of instances already registered for each of the SW, these are called syntactic word locations (SWL). In the case of $\frac{1}{2}$ / STUDY / nab, that is an abstract noun, there are 41 text locations, some of which can be seen in Figure 9. Here again the text is given with the translation and there is also a link, which allows to inspect the text location in its context. Figure 10 gives the list of SWL for $\frac{1}{2}$ / STUDY / vt[oN], that is as a transitive verb with a noun object.

4 Use cases for the new tagset

Before moving to a more detailed description of the new tagset, a list of use cases the system hopes to address might be useful. This list only gives desirables of functionality not currently available.

4.1 Describe multiple versions of texts

Multiple versions in different text formats have to be representable, they should be accessible simultaneously to facitilitate text comparison.

It should be easy and straightforward to add new text versions to the collection.

Special cases also need consideration, such as cases where the order of the text is different in some versions.

1 《學而篇第一》	Translation by C. Harbsmeier *	學 (xué) S • nab a	TUDY New SW	SYN 20 learn about things (typica	ally from a
1 《学面篇第一》 line 1 / 0% 1 《學而篇第一》 1.1子曰: 「四而時習之, 不亦說乎? 有朋自遠方來, 不亦樂乎? 人不知而不慍, 不亦君子乎?」 1.2有子曰:	Translation by C. Harbsmeier • The Master said, "Having studied something to exercise it in practice, at the proper times, isn't that satisfying? To have a colleague come from a distant region isn't that delightful? When others do not appreciate one not to feel offended isn't that behaving in a gentlemanly fashion?" Yôuzi said,	 nab a nab a teacher) str learning id記 43 荀子 1 荀子 1 荀子 1 荀子 1 	ct: the attempt to Jdy; the pursuit o iawe Use sw 43.1大學之 道 學不可以 已。 學惡乎始? 學至乎沒而 後止也,	f intellectual/moral self-de 1.: 41 1. What the Great Lea "In studying one must Where should study be Study first ends with o	velopment; rning teaches, کیر not stop." کیر egin? کیر ne's death. یر ک
「其為人也孝弟,而好犯上 者, 鮮矣; 不好犯上,而好作亂者, 未之有也。 君子務本, 本立而道生。	"That anyone whose personality is both filial and fraternal and yet should be prone to offend the authorities, that is quite rare. For someone not prone to offending superiors yet to be prone to creating political rebellions, that never occurs. The gentleman concentrates his efforts on what is basic. Once what is basic is established, then the Way will [naturally] emerge.	荀子 1 荀子 1 荀子 1 荀子 1	故學數有 終, 故學至乎 《禮》而止 矣。 小人之學 也: 學更乎近	Thus the art of study h point, Thus, when study gets of ritual, then it stops. As for the small man's In study nothing is more	as a final ightarrow imes to the point ightarrow imes studying: $ ightarrow imes$ re effective
孝弟也者, 其為人之本與!」 1.3子曰: 「巧言令色, 鮮矣仁!」	As for filial piety and fraternal love (in general), these must count as the basis for Goodness, mustn't they?" The Master said, "Those who make their speech crafty and their appearance artificially distinguished are rarely indeed Good."	荀子 1 荀子 2	其人。 學莫便乎近 其人。 故學也者,	than being close to the In study nothing is mo than be close to the rig Therefore one who is i of learning	e right person. re effective ght person. mthe process

Figure 9: Text references for STUDY/ $\ensuremath{\mathbb{P}}$ xué nab (excerpt)

		 vt[oN] act: devote oneself to study; be devoted to study; engage 			
1《学而扁弟一》	Translation by C. Harbsmeier 🝷	in intelle	ectual work; work to i	improve oneself morally/intellectually	
line 1 / 0%		Save	Use SWL: 34		
1 《學而篇第一》		禮記:	禮聞來學,	I have heard in the same way of	
1.1子曰:	The Master said,			(scholars) coming to learn; ∆x	
「 <mark>學</mark> 而時習之,	"Having studied something to exercise it in practice, at the proper times,	荀子:	1 君子博學	If the gentleman will study broadly 🗙	
不亦說乎?	isn't that satisfying?			ப்	
有朋自遠方來,	To have a colleague come from a distant region	荀子:	L 不足謂善	then he does not deserve to be	
不亦樂乎?	isn't that delightful?		學。	called good at studying. 👌	
人不知而不慍,	When others do not appreciate one not to feel offended	荀子:	2 加好學遜敏	If you add to these a love of learning	
不亦君子乎?」	isn't that behaving in a gentlemanly fashion?"		焉,	ن لا	
1.2有子曰:	Yŏuzi said,	列子:	B 同師而學,	studied under the same teachers, $ {f x}$	
「其為人也孝弟,而好犯上	"That anyone whose personality is both filial and fraternal and yet should			ථ	
者,	be prone to offend the authorities,	說苑:	3 少而好學,	۵×	
鮮矣;	that is quite rare.	說苑:	3 老而不學,	but if as an old age one does not	
不好犯上,而好作亂者,	For someone not prone to offending superiors yet to be prone to creating			devote oneself to study	
	political rebellions,	說苑:	L 「寡人願學	۵×	
未之有也。	that never occurs.		而無師。」		
君子務本,	The gentleman concentrates his efforts on what is basic.	孟子:	3 我學不厭而	I study without getting fed up and	
本立而道生。	Once what is basic is established, then the Way will [naturally] emerge.		教不倦	teach without getting tired'	
孝弟也者,	As for filial piety and fraternal love (in general),		也。』		
其為人之本與!」	these must count as the basis for Goodness, mustn't they?"	孟子:	3 『學不厭,	'To study without getting fed up	
1.3子曰:	The Master said,	法言:	L 學以治之,	THE PURPOSE OF STUDY IS TO	
「巧言令色,	"Those who make their speech crafty and their appearance artificially			sort oneself out;	
	distinguished	論語:	L 「學而時習	"Having studied something to	
鮮矣仁!」	are rarely indeed Good."		之,	exercise it in practice, at the proper	
or Direfox Web Brows				🔼 🔿 🕶 in 🚽 📿 📣 i	

Figure 10: Text references for STUDY/學 xué vt[oN] (excerpt)

4.2 Collection of fragmentary texts / anthology

Virtual description of fragments of texts should be possible, by describing formally what text span from where has to be taken. This could also be an anthology of works in different collections.

4.3 Text and commentary

Linking of root texts and their commentaries: Looking at a text line, all relevant commentaries should be retrievable.

4.4 Annotation, translation, markup

In addition to describing and interlinking existing textual sources, it also should be possible to create and describe new content such as annotations, translations and marking of textual features.

4.5 Groups of related, but distinct texts

Some texts cover related topics, without having a strong textual dependency. To accomodate the reasoning about such texts, a mechanism for dealing with groups of related texts is necessary as a higher level of description.

4.6 Analytic tools

Making the texts available in a simple, standardized form makes it easier to develop analytical tools or to interface with existing tools.

5 XML formats for supporting the infrastructure

Rather than converting the texts in the Kanseki Repository to XML, we continue to maintain the texts using the existing format and tools and connect to the other components of the research environment using a number of predefined XML formats. This has the additional advantage of providing an open architecture with the potential of other applications connecting to the research environment or using parts of it through the same API interfaces. It will also allow to interact with texts in other formats, for example the texts in the TLS, which are currently using TEI XML⁵.

Herer are some functional requirements, that need to be catered for:

- grouping and selecting of texts (manifest)
 - this also includes identifying and relating parts of texts in a regular way, such as a root text to commentaries or translations.
 - It will also be a convenient way to maintain metadata related to the texts.
 - The implementation should allow users to create their own manifest files, which will make it possible to make collections and metadata adjustable.

⁵An XML format defined by the Text Encoding Initiative (TEI), see https://www.tei-c.org.

- links between text passages (nexus)
- a format that can be processed to directly compare texts algorithmically (token)

5.1 Practical considerations

The manifest provides the information the different parts of the system need to interoperate. This includes for example the information where a certain version of a text is located (in the case of the KR, this will include the text ID number and the branch) and what format this version is in.

A utility program (or a API interface) will use this information to create a token file. This file contains one <t> element for each character in the edition, optionally grouped in a potentially nested hierarchy of <tg>, i.e. token group element. For examples of these usages please see the tagset documentation provided in the Appendix. If there are several editions representing one text, such a token file will be prepared for each of the editions.

Based on these token files, another utility program can be used to produce the **nexus** files, we currently maintain one such file for every edition, thus providing a view from that specific edition to all the other editions involved. In the interface of the collaborative research platform, this will provide a means to call up different versions of a phrase unter investigation, or provide quick access to the location of this phrase in commentaries. A practical example again is provided in the tagset documentation.

6 KanripoX preview

As a preview of how the information contained in the KanripoX XML files can be used to display additional versions of a text displayed, Figure 11 shows a mockup of this information for a specific line of text. This is still very preliminary and is bound to change. As can be seen, both information on alternate text version, as well as commentaries to the line in question are available.

The token files can also be produced from the various text formats by scripts available in the KanripoX GitHub repository. Based on these files, other scripts are used to write the nexus files, that list the corresponding text passages. The format is also compatible the CollateX program, which can be used to compare multiple text variants and visualize the results.

7 Conclusion

The proposed direction of development for a new collaborative research platform is still a rough sketch and will need further refinement as the platform is developed and used in practice. It is hoped that the XML format developed here can be used to support this and provide the missing link between specific editions that are maintained in different formats.

學 삹火 TLS Browse ▼ :	老子 目錄 ▼ 🗿 Source: CHANT Bookmarks ▼ 內部 ▼	Search in texts Q & chris -
1 第一章 line 8 / 57%	Translation by Ursula K. Le Guin 🕶	○ 老子 (CH1a0918a_chant) 故恒無欲,以觀其妙。
1 第一章	Taoing {A satisfactory translation of this chapter is, I believe, perfectly impossible. It contains the book. I think of it as the Aleph, in Borges's story: if you see it rightly, it contains everything.}	 老子 (CH1a0918a_other) 故▶(常) ◀(m:編者按:《馬王堆漢墓帛書甲本老子》頁19作 「恒无欲也」,今本作「常」者,蓋避漢文帝諱改。▶(恒)
道可道,	The way you can go	【fn::編者按:《馬王堆漢墓帛書甲本老子》頁19作「恒无欲也」,今本作「常」者,蓋避漢文帝諱改。]無欲◀fn::馬王堆漢
非恒道。	isn't the real way.	墓帛書甲本老子頁19],以觀其妙。常)◀[fn::編者按:《馬王堆
名可名,	The name you can say	漢墓帛書甲本老子》頁19作「恒有欲也」,今本作「常」者,蓋
非恒名。	isn't the real name.	避漢文帝諱改。]▶〔
無名、天地之始,	Heaven and earth begin in the unnamed:	○ 老子(河上公注) (CH1a0918b_chant)
有名、萬物之母,	name's the mother of the ten thousand things.	故常無欲,以觀其妙。
故恒無欲,以觀其妙。	So the unwanting soul sees what's hidden,	 老子(河上公注) (CH1a0918b_other) 故常無欲,以觀其妙。
恒有欲,以觀其徼。	and the ever-wanting soul sees only what it wants.	○ 老子甲本 (CH8x3004_chant)
此兩者同出而異名。	Two things, one origin, but different in name,	故恒无欲也,以觀其眇;
同調之女。	whose identity is mystery	◎ 老子甲本卷後古佚書 (CH8x3005_chant)
玄之又玄,	Mystery of all mysteries	是則,以肥天地。禮數
眾妙之門。	The door to the hidden.	○ 老子乙本 (CH8x3007_chant) 故恒无欲也,以觀其妙

Figure 11: First chapter of Laozi with alternate editions displayed for line "故恒無欲,以 觀其妙"

8 References

- Burnard, Lou, and Syd Bauman, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Oxford, Providence, Charlottesville, Nancy: TEI Consortium. 2008.
- Middell, Gregor and Ronald Haentjens Dekker, *CollateX Software for Collating Textual Sources* [accessed 2021-02-12], https://collatex.net/.
- Harbsmeier, Christoph, "Philological Reflections on Chinese Conceptual History: Introducing Thesaurus Linguae Sericae", in: *Keywords in Chinese Culture*, ed. by Wai-yee Li and Yuri Pines, The Chinese University of Hong Kong Press, 2020, p. 381-403.
- KanripoX, [accessed 2021-02-12], https://github.com/kanripox/kanripox-dev.
- Wittern, Christian (ed.), *Special issue Kanseki Repository*, CIEAS Research Report 2015, Kyoto 2016 available at http://hdl.handle.net/2433/210140.

Appendix: Schema XML files in the KanripoX project

1 Overview

The following sections detail the different file formats that have been defined for the extension of the Kanseki Repository. Although they constitute very different information for different purposes, for the convenience of describing the files and processing the information, they have been combined into one single schema, adressed under one single namespace. The schema allows the following entry points:

<manifests> Root for manifests that contain multiple manifest elements.

 $<\!manifest\!>$ The root of the manifest. One manifest describes one work.

<tList> Root for token that may contain one or more **<**tg> elements.

<nexusList> Root for Nexus that may contain one or more <nexus> elements.

The first two of these pertain to the manifest files; the <manifest> definitions can be grouped together into a list of <manifests>, thus providing two entry points for the schema for manifests, while the other two schemas define a list of <nexus> respectively <t> (token) elements, which are grouped together into lists, thus the lists provide the entry point in this case.

2 Grouping and description of texts: the manifest

The Manifest.xml described here contains information about a set of editions that are grouped here together, usually for the purpose of further description and processing.

There are two main elements under the root element <manifest $>^1$:

<editions> The editions representing the work under consideration. Work is taken in a very broad sense here.

divisions The internal subdivisions of the work under consideration.

The <editions> element holds information about the editions that are collected here. It may contain either <editionGroup> elements (which hold <edition> elements), or directly contain <edition> elements, which give the details for each edition. This includes also the *type*, which will be either "documentary" or "interpretative". Documentary editions are editions that strive to reproduce an existing print edition, while interpretative editions do reflect the views of the editor and do not follow one single edition.

Other details for editions that will be collected here are the *id*, which is a unique label (or identifier) used to refer to this specific edition within the manifest and the processing systems.

The <edition> element may have the following children:

<title> Title of the work.

<creation> Information about the creation: date and responsible agent.

<description> Description of the edition or item this element is attached to.

<tokenmap> Mappings from textual features to token types.

<divisions> The internal subdivisions of the work under consideration.

All of these elements are optional. It is useful to give a <title>, which will make references to this edition easier. The <creation> element gives role, function and date of the persons involved in creating this edition. This should refer to the specific edition, more general information pertaining to all editions in this group can be given in the <editionGroup>. The <description> contains other relevant information concerning the edition. <tokenmap> is a way to map features of the text to specific token attributes, details of this are given below. And finally, the <divisions> allows reference to divisions within the edition. This will record the divisions of this specific edition, a more abstract view of editions for the whole group of texts can be given

¹There are in fact two possible root elements, the other being <manifests> for a grouping of manifest elements.

as a direct child to <manifest>. If the order of the division is not the same in all divisions, they can be re-arranged here and linked to the text location.

Here is an example for an annotated version of the Daode jing:

```
<editionGroup type="annotation">
<editionformat="xml/TEI"
id="KX5c0065_HFL" language="lzh"
location="doc/KX5c0065_HFL" type="documentary">
<creation>
<date notafter="-100" notbefore="-200">1st century CE</date>
<resp role=" 注"> 河上公 </resp>
</creation>
<description> 道德真經 </description>
</edition>
```

The $\langle \text{divisions} \rangle$ element can also occur as a child element of manifest, optionally following the $\langle \text{editions} \rangle$ element. If used here, there will be only one division element, which holds all subdivisions as possibly nesting $\langle \text{div} \rangle$ elements. The purpose of this element is to provide an entry point to the editions, which is neither tied to one specific edition, nor to a hyperlink or similar in a technical sense. The *label* on the $\langle \text{div} \rangle$ elements is used to provide a human readable label that can be used to point to that specific division, much the same as "Chapter 2" will (usually) refer to the same section of a work, no matter which edition is used. To serve as a link between this nesting structure of chapters, sections and so forth, each $\langle \text{div} \rangle$ can have one or more $\langle \text{edRef} \rangle$ elements, which point to the text span in one of the editions that is covering this specific section.

```
<div label=" 第一章">
<edRef end="61" key="KR5c0057_tls"
start="0"/>
<edRef end="58" key="CH1a0918a_chant"
start="0"/>
<edRef end="402" key="CH1a0918b_chant"
start="2"/>
</div>
```

In this example, the *start* and *end* attributes give the number of the first and last token that is part of this section of a text, thus identifying the text span independent of the text format of the text. Other possibilities for adressing a text span are available if the edition is in TEI/XML.

3 A shadow of the text: the token file

The token files described here serve as a shadow of other digital files that more thoroughly describe the texts documented there. This relieves the token files from the burden to describe the physical appearence, structure and transmission of the text. This information is available at any time by following the links back to these other files. The purpose of the token files is to provide a minimal description, containing only the characters of the text in a form that allows easy comparison and alignment of multiple versions and a very minimal hierarchical structure if necessary. The function is similar to a concordance in that it provides access to the whole text, but without much of what a reader would expect to make reading (or editing) convenient, or even feasible. On the other hand, enough information should be retained to reconstruct a very basic version of the text.

The main elements under the root element <tList>are:: <tg> A group of tokens. <t> A token. The $\langle tg \rangle$ element holds the $\langle t \rangle$ elements, which have the character content of the text, one token per $\langle t \rangle$. The purpose of the $\langle tg \rangle$ element is to group related $\langle t \rangle$ elements. $\langle tg \rangle$ can nest, and provide thus for a rudimentary structure in the token files.

Here is an example of a token file, for the beginning of the Daode jing by Laozi:

```
<tList xmlns="http://kanripo.org/ns/KRX/1.0" ed="KR5c0057_tls" n="tok-0000"
xml:id="KR5c0057_tls-tok-0000">
<tg>
 <t n="KR5c0057_tls_001-1a.3-h" pos="1"
  role="h" tp="0"> 第 </t>
 <t n="KR5c0057_tls_001-1a.3-h" pos="2"
  role="h" tp="1"> - </t>
 <t n="KR5c0057_tls_001-1a.3-h" pos="3"
  role="h" tp="2"> 章 </t>
</tg>
<tg xml:id="KR5c0057_tls_001-1a.3">
 <lb ed="KR5c0057_tls"
  n="KR5c0057_tls_001-1a.3"/>
 <t n="KR5c0057_tls_001-1a.3" pos="1"
  role="p" tp="3"> 道 </t>
 <t n="KR5c0057_tls_001-1a.3" pos="2"
  role="p" tp="4"> 可 </t>
 <t f="," n="KR5c0057_tls_001-1a.3" pos="3"
  role="p" tp="5"> 道 </t>
</tg>
<tg xml:id="KR5c0057_tls_001-1a.4">
 <lb ed="KR5c0057_tls"
  n="KR5c0057_tls_001-1a.4"/>
 <t n="KR5c0057_tls_001-1a.4" pos="1"
  role="p" tp="6"> 非 </t>
 <t n="KR5c0057_tls_001-1a.4" pos="2"
  role="p" tp="7"> 恒 </t>
 <t f=" " n="KR5c0057_tls_001-1a.4" pos="3"
  role="p" tp="8"> 道 </t>
</tg>
<tg xml:id="KR5c0057_tls_001-1a.5">
 <lb ed="KR5c0057_tls"
  n="KR5c0057_tls_001-1a.5"/>
 <t n="KR5c0057_tls_001-1a.5" pos="1"
  role="p" tp="9"> 名 </t>
 <t n="KR5c0057_tls_001-1a.5" pos="2"
  role="p" tp="10"> 可 </t>
 <t f="," n="KR5c0057_tls_001-1a.5" pos="3"
  role="p" tp="11"> 名 </t>
</tg>
<tg xml:id="KR5c0057_tls_001-1a.6">
 <lb ed="KR5c0057_tls"
  n="KR5c0057_tls_001-1a.6"/>
 <t n="KR5c0057_tls_001-1a.6" pos="1"
 role="p" tp="12"> 非 </t>
 <t n="KR5c0057_tls_001-1a.6" pos="2"
  role="p" tp="13"> 恒 </t>
 <t f="°" n="KR5c0057_tls_001-1a.6" pos="3"
  role="p" tp="14"> 名 </t>
</tg>
<tg xml:id="KR5c0057_tls_001-1a.7">
 <lb ed="KR5c0057_tls"
  n="KR5c0057_tls_001-1a.7"/>
 <t n="KR5c0057_tls_001-1a.7" pos="1"
  role="p" tp="15"> 無 </t>
 <t f="`` n="KR5c0057_tls_001-1a.7" pos="2"
```

```
role="p" tp="16"> 名 </t>
  <t n="KR5c0057_tls_001-1a.7" pos="3"
   role="p" tp="17"> 天 </t>
  <t n="KR5c0057_tls_001-1a.7" pos="4"
   role="p" tp="18"> 地 </t>
  <t n="KR5c0057_tls_001-1a.7" pos="5"
   role="p" tp="19"> 之 </t>
  <t f="," n="KR5c0057_tls_001-1a.7" pos="6"
   role="p" tp="20"> 始 </t>
 </tg>
 <tg xml:id="KR5c0057_tls_001-1a.8">
  <lb ed="KR5c0057_tls"
   n="KR5c0057_tls_001-1a.8"/>
  <t n="KR5c0057_tls_001-1a.8" pos="1"
   role="p" tp="21"> 有 </t>
  <t f=" \" n="KR5c0057_tls_001-1a.8" pos="2"
   role="p" tp="22"> 名 </t>
  <t n="KR5c0057_tls_001-1a.8" pos="3"
   role="p" tp="23"> 萬 </t>
  <t n="KR5c0057_tls_001-1a.8" pos="4"
   role="p" tp="24"> 物 </t>
  <t n="KR5c0057_tls_001-1a.8" pos="5"
   role="p" tp="25"> 之 </t>
  <t f="," n="KR5c0057_tls_001-1a.8" pos="6"
   role="p" tp="26"> ↔ </t>
 </tg>
<!-- many more tokens ... -->
</tList>
```

4 Links between text passages: the nexus file

The nexus files described here describe links between locations in texts. The links consist of references to a span of one or more consecutive characters in a text, machine readably expressed in terms of references to the $\langle t \rangle$ elements in the token files. Related links can be grouped together to form a nexus. This can be used for example to describe corresponding passages in different versions of a text.

The main elements under the root element <nexusList>are::

<nexus> A group of **<**locationRef**>** elements.

<note> An additional note.

The <nexus> element holds the <locationRef> elements, which contain the reference information to locate the passage of the text. The reference is expressed by pointing to a sequence of one or more tokens in a token file for the edition.

This example shows a <nexus> element from a nexus file for the edition with the identifier "KR5c0057_tls" in the manuscript file. the *tcount* tells us that 6 tokens are involved in this edition, the number of tokens in the other editions is given in an attribute of the same name on the <locationRef> elements. Some have longer sequences, in these cases the commentary pertaining to this phrase has also been included in the parallel text. The *tp* gives the token position, that is the sequential number of the token in the edition, from here it can be seen that in some editions, the corresponding text is not at the beginning, either because of a different sequence (as in "CH8x3004_chant" and "CH8x3007_chant") or because a preface occurs before the text proper, as in "KX5c0065_SBCK".

```
<nexus tcount="6" tp="15"

xml:id="KR5c0057_tls_001-1a.7">

<locationRef ed="CH1a0918a_chant"

target="CH1a0918_CHANT_001-1a.6" tcount="6" tp="14"/>
```

```
<locationRef ed="CH1a0918a_other"
  target="CH1a0918_CHANT_001-1a.6" tcount="7" tp="26"/>
 <locationRef ed="CH1a0918b_chant"</pre>
  target="CH1a0918_CHANT_082-4a.3" tcount="42" tp="121"/>
 <locationRef ed="CH1a0918b_other"
  target="CH1a0918_CHANT_082-4a.3" tcount="41" tp="119"/>
 <locationRef ed="CH8x3004_chant"
  target="CH8x3004_CHANT_002-1a.15" tcount="7" tp="3063"/>
 <locationRef ed="CH8x3005_chant"
  target="CH8x3005_CHANT_002-1a.44" tcount="6" tp="5072"/>
 <locationRef ed="CH8x3006_chant"
  target="CH8x3006_CHANT_001-1a.192" tcount="6" tp="389"/>
 <locationRef ed="CH8x3007_chant"</pre>
  target="CH8x3007_CHANT_002-1a.19" tcount="7" tp="3060"/>
 <locationRef ed="KR5c0073_tls"
  target="KR5c0073_tls.seg2-1" tcount="6" tp="35"/>
 <locationRef ed="KX5c0045_HFL"
  target="KX5c0045_HFL_001-001a.03" tcount="6" tp="24"/>
 <locationRef ed="KX5c0045_ZTDZ"
  target="KX5c0045_SJB_001-110474b.03" tcount="6" tp="24"/>
 <locationRef ed="KX5c0046_HFL"
  target="KX5c0046_HFL_000-001a.03" tcount="6" tp="27"/>
 <locationRef ed="KX5c0046_ZTDZ"</li>
  target="KX5c0046_SJB_000-110482a.03" tcount="6" tp="27"/>
 <locationRef ed="KX5c0065_SBCK"
  target="KX5c0065_SBCK_001-1a.06" tcount="38" tp="1020"/>
 <locationRef ed="KX5c0065_ZTDZ"
  target="KX5c0065_SJB_001-120001a.10" tcount="40" tp="119"/>
 <locationRef ed="KX5c0073_HFL"
  target="KX5c0073_HFL_001-001a.05" tcount="6" tp="50"/>
 <locationRef ed="KX5c0073_ZTDZ"
  target="KX5c0073_SJB_001-120272c.05" tcount="6" tp="50"/>
</nexus>
```

5 Schema for Manifest, Token and Nexus

5.1 Elements

<creation> Information about the creation: date and responsible agent.

Module KRXManifest Contained by description edition editionGroup May contain KRXManifest: date resp Content model

```
<content>
<alternate maxOccurs="unbounded"
minOccurs="0">
<elementRef key="date"/>
<elementRef key="resp"/>
</alternate>
</content>
```

Schema Declaration

element creation { (krx_date | krx_resp)* }

<date> Date of the work.

Module KRXManifest Attributes **Onotbefore** Earliest possible date. Status Optional Datatype string Onotafter Latest possible date. Status Optional Datatype string **@cert** Degree of certainty of this assertion. Status Optional Legal values are: high High degree of certainty. middle Middle degree of certainty. low Low degree of certainty.

Contained by creation

May contain Character data only

Content model <content> <textNode/></content>

Schema Declaration

```
element date
ſ
  attribute notbefore { text }?,
  attribute notafter { text }?,
  attribute cert { "high" | "middle" | "low" }?,
   text
3
```

 $<\!\! description\! >$ Description of the edition or item this element is attached to.

Module KRXManifest Contained by div edition manifest May contain *KRXManifest:* creation note title character data

Content model

```
<content>
<alternate maxOccurs="unbounded"</pre>
 minOccurs="0">
 <textNode/>
 <elementRef key="note" minOccurs="0"/>
 <elementRef key="title" minOccurs="0"/>
 <elementRef key="creation" minOccurs="0"/>
</alternate>
</content>
```

Schema Declaration

element description { (text | krx_note? | krx_title? | krx_creation?)* }

<div**>** One specific subdivision on any level.

Module KRXManifest

Attributes

@label A label to identify the subdivision, can be any string, but should be unique in the manifest. This can be used to access this textual division. *Status* Optional

Datatype token

 ${\tt @edition}$ A reference to the edition, as defined elsewhere in this manifest. Status Optional

Datatype IDREF

Osequence Sequencial number of this division, given in such a way that ordering by this number will produce the text in the same sequence as the base edition.

Status Optional

Datatype nonNegativeInteger

 ${\tt @start}$ The sequencial number of the first token of this division in the token list. Status Optional

Datatype nonNegativeInteger

 ${\tt @end}$ The sequencial number of the last token of this division in the token list. Status Optional

Datatype nonNegativeInteger

@divid If the source file of this edition has an identifier (usually a xml:id for this subdivision), it can be recorded here.

Status Optional

Datatype token

Contained by div divisions

May contain

KRXManifest: description div edRef label

Content model

<content>

```
<sequence maxOccurs="1" minOccurs="1">
  <elementRef key="label"
  maxOccurs="unbounded" minOccurs="0"/>
  <elementRef key="description"
  minOccurs="0"/>
  <elementRef key="edRef"
  maxOccurs="unbounded" minOccurs="0"/>
  <elementRef key="div"
  maxOccurs="unbounded" minOccurs="0"/>
  </sequence>
</content>
```

Schema Declaration

```
element div
{
    attribute label { text }?,
    attribute edition { text }?,
    attribute sequence { text }?,
```

```
attribute start { text }?,
attribute end { text }?,
attribute divid { text }?,
( krx_label*, krx_description?, krx_edRef*, krx_div* )
}
```

<divisions> The internal subdivisions of the work under consideration.

Module KRXManifest

Attributes

@edition If necessary, the edition for which these textual divisions are valid can be given here. Status Optional Datatype token

Contained by edition manifest May contain KRXManifest: div

Content model

```
<content>
<elementRef key="div"
maxOccurs="unbounded" minOccurs="1"/>
</content>
```

Schema Declaration

element divisions { attribute edition { text }?, krx_div+ }

<edRef> Reference to this subdivision in one specific edition, identified by the key.

Module KRXManifest

Attributes

Ostart The sequencial number of the first token of this division in the token list. Status Optional

Datatype nonNegativeInteger

 ${\tt @end}$ The sequencial number of the last token of this division in the token list. Status Optional

Datatype nonNegativeInteger

Okey A reference to the edition, as defined elsewhere in this manifest.

Status Optional

Datatype IDREF

@timestamp The timestamp in ISO format, e.g. 2020-10-09T14:23:52+09:00. *Status* Optional

Datatype dateTime

@label A label to identify the subdivision as used in this edition. It can be any string, but should be unique in the manifest. This can be used to access this textual division. Status Optional Datatype token

Contained by div May contain Empty element Content model <content> <empty/></content> Schema Declaration

```
element edRef
{
   attribute start { text }?,
   attribute end { text }?,
   attribute key { text }?,
   attribute timestamp { text }?,
   attribute label { text }?,
   empty
}
```

<edition> One edition of the work. If there are multiple <divisions>, this indicates the sequence of these divisions make up the work.

Module KRXManifest

Attributes

@xml:id The identifier of the work. This will be used to refer to this manifest from the display of this text.

Status Optional

Datatype ID

©id The identifier of the edition. This is required and has to be unique within this manifest. It will be used by the processing tools to refer to this edition. *Status* Required

Datatype ID

@format The parsing tool is selected based on the format given here, there are two formats defined at the moment. Additional formats can be added, but require a plugin to parse them.

Status Required

Legal values are: xml/TEI TEI file encoded in XML.

txt/mandoku Mandoku format.

Clocation This gives either the relative path to the local folder containing the edition or a resolvable remote reference to the edition, for example on github.

Status Required

Datatype string

Note TODO: format for remote reference.

TODO: Format for identifying portion of text in file.

@base The edition marked as 'base' is the reference edition for sequential reordering.

Status Optional

Legal values are: true This edition is the reference edition.

false This edition is not the reference edition (default).

 $\ensuremath{\texttt{Otype}}$ The edition has to be declared as either 'documentary' or 'interpretative' .

Status Required

- Legal values are: **documentary** An edition that documents an existing print source as faithful as possible, without editorial changes.
 - **interpretative** An edition that might be based on a print source, but possibly makes editorial changes.
- **Orole** One of the editions has to be declared as the base edition, the others are reference editions.

Status Recommended

Legal values are: base This edition is the base edition.

reference All editions except the base edition are considered reference editions.[Default]

@language The language of the document, identified with an identifier according to RFC 1766.

Status Optional

Datatype language

Osigle A short identifier used to identify this edition.

Status Optional

Datatype string

Contained by editionGroup editions

May contain

KRXManifest: creation description divisions title tokenmap

Content model

```
<content>
<sequence maxOccurs="1" minOccurs="1">
<elementRef key="title" maxOccurs="1"
minOccurs="0"/>
<elementRef key="creation" maxOccurs="1"
minOccurs="0"/>
<elementRef key="description"/>
<elementRef key="tokenmap" maxOccurs="1"
minOccurs="0"/>
<elementRef key="divisions"
maxOccurs="unbounded" minOccurs="0"/>
</sequence>
</content>
```

Schema Declaration

```
element edition
{
  attribute xml:id { text }?,
  attribute id { text },
  attribute format { "xml/TEI" | "txt/mandoku" },
  attribute location { text },
  attribute base { "true" | "false" }?,
  attribute type { "documentary" | "interpretative" },
  attribute role { "base" | "reference" }?,
  attribute language { text }?,
  attribute sigle { text }?,
   (
      krx_title?,
     krx_creation?,
      krx_description,
      krx_tokenmap?,
```

<editionGroup> A group of the editions representing the work under consideration.

Module KRXManifest

Attributes

@type The treatment of the editions within this group are based on the value of this attribute. *Status* Required

Legal values are: root The root text of this work.

root+annotation The root text, interspersed with commentary.

annotation Commentary to the root text, without repeating the text.

translation Translations of the text and / or commentary.

other Texts, that are grouped with this texts for some reason other than being textually related.

Osigle A short identifier used to identify this group of editions.

Status Optional

Datatype string

Contained by editions

May contain KRXManifest: creation edition title Content model

```
<content>
<sequence maxOccurs="1" minOccurs="1">
<elementRef key="title" maxOccurs="1"
minOccurs="0"/>
<elementRef key="creation" maxOccurs="1"
minOccurs="0"/>
<elementRef key="edition"
maxOccurs="unbounded" minOccurs="1"/>
</sequence>
</content>
```

$Schema \ Declaration$

```
element editionGroup
{
    attribute type
    {
        "root" | "root+annotation" | "annotation" | "translation" | "other"
    },
    attribute sigle { text }?,
    ( krx_title?, krx_creation?, krx_edition+ )
}
```

<editions> The editions representing the work under consideration. Work is taken in a very broad sense here.

Module KRXManifest Contained by manifest May contain KRXManifest: edition editionGroup Content model

```
<content>
<alternate maxOccurs="1" minOccurs="1">
<elementRef key="editionGroup"
maxOccurs="unbounded" minOccurs="1"/>
<elementRef key="edition"
maxOccurs="unbounded" minOccurs="1"/>
</alternate>
</content>
```

 $Schema \ Declaration$

element editions { krx_editionGroup+ | krx_edition+ }

<label> Additional label.

Module KRXManifest
Attributes
Clanguage The language of the label, identified with an identifier according to RFC 1766.
Status Optional
Datatype language
Contained by div
May contain Character data only

Content model <content> <textNode/></content>

Schema Declaration element label attribute language text ?, text

This element marks the beginning of a new line or line-like section on the text-bearing surface.

Module derived-module-KRX

Attributes

@ed Identifier of the edition to which this line belongs. Status Optional Datatype string
@n Number or other label used to refer to this line. Status Optional Datatype string
@xml:id Status Recommended

amina Status Recommence

Datatype ID

KRXToken: tg

May contain Empty element

Content model <content> <empty/></content>

 $Schema \ Declaration$

```
element lb
{
    attribute ed { text }?,
    attribute n { text }?,
    attribute xml:id { text }?,
    empty
}
```

Reference to a location in the token file. Optionally might hold a copy of the referenced text as a string of characters.

Module KRXNexus

Attributes

Oed Identifier of the edition (as used in the token file).

Status Required

Datatype string

@tp The sequencial number of the first token in the token file. *Status* Required

Datatype nonNegativeInteger

Otcount The number of tokens that make up this text span.

Status Optional

Datatype nonNegativeInteger

Default 1

@target Identifier of the first token in the text span.

Status Required

Datatype string

On Label or identifier for this reference. *Status* Optional

Datatype string

KRXNexus: nexus

May contain Character data only Content model <content> <textNode/></content> Schema Declaration

```
element locationRef
{
    attribute ed { text },
    attribute tp { text },
    attribute tcount { text }?,
    attribute target { text },
    attribute n { text }?,
    text
}
```

<manifest> The root of the manifest. One manifest describes one work.

Module KRXManifest

Attributes

@xml:id The identifier of the work. This will be used to refer to this manifest
from the display of this text.
Status Optional
Datatype ID

Contained by manifests

May contain

KRXManifest: description divisions editions title

Note Currently, only one work can be described per one manifest file. Need to think about what to do with use cases that need multiple works. Use several <manifest> in a file?

Content model

<content>

```
<sequence maxOccurs="1" minOccurs="1">
  <elementRef key="title" minOccurs="0"/>
  <elementRef key="description"/>
  <elementRef key="divisions"/>
  <elementRef key="divisions" minOccurs="0"/>
  </sequence>
</content>
```

Schema Declaration

```
element manifest
{
    attribute xml:id { text }?,
    ( krx_title?, krx_description, krx_editions, krx_divisions? )
}
```

<manifests> Root for manifests that contain multiple manifest elements.

Module KRXManifest —

May contain KRXManifest: manifest Content model

```
<content>
<elementRef key="manifest"
maxOccurs="unbounded"/>
</content>
```

Schema Declaration | element manifests krx_manifest+

<map> Map of one textual feature to a specific token type.

Module derived-module-KRX

Attributes

Osrc Element or simple matching expression (for XML texts) or regular expressions (for plain text) that identifies the textual feature.

Status Optional Datatype string ©tok Token type. Status Optional Legal values are: h Token is part of a heading. p Token is part of a paragraph. n Token is part of a note or annotation of any kind. q Token is part of a quotation. v Token is part of a verse line Contained by tokenmap

May contain Empty element

```
Content model <content> <empty/></content>
```

Schema Declaration

```
element map
{
    attribute src { text }?,
    attribute tok { "h" | "p" | "n" | "q" | "v" }?,
    empty
}
```

<nexus> A group of <locationRef> elements.

Module KRXNexus
Attributes **Q**xml:id The identifier of this token group.
Status Optional
Datatype ID **Qtp** The sequencial number of the first token of this text span.
Status Required
Datatype nonNegativeInteger

@tcount The number of tokens that make up this text span.

Status Optional Datatype nonNegativeInteger Default 1

KRXNexus: nexusList

May contain

KRXManifest: note

KRXNexus: locationRef

 $Content\ model$

<content>

```
<sequence maxOccurs="1" minOccurs="1">
  <elementRef key="note"
   maxOccurs="unbounded" minOccurs="0"/>
   <elementRef key="locationRef"
   maxOccurs="unbounded" minOccurs="0"/>
   </sequence>
</content>
```

Schema Declaration

```
element nexus
{
    attribute xml:id { text }?,
    attribute tp { text },
    attribute tcount { text }?,
    ( krx_note*, krx_locationRef* )
}
```

<nexusList> Root for Nexus that may contain one or more <nexus> elements.

Module KRXNexus Attributes ©xml:id Status Recommended Datatype ID ©ed Reference to the edition defined in the manifest. Status Required Datatype string ©n A label Status Optional Datatype string

May contain KRXManifest: note KRXNexus: nexus Content model

```
<content>
<sequence maxOccurs="1" minOccurs="1">
        <elementRef key="note" maxOccurs="1"
        minOccurs="0"/>
        <elementRef key="nexus"
        maxOccurs="unbounded"/>
        </sequence>
</content>
```

Schema Declaration

```
element nexusList
{
    attribute xml:id { text }?,
    attribute ed { text },
    attribute n { text }?,
    ( krx_note?, krx_nexus+ )
}
```

<note> An additional note. Module KRXManifest Contained by description KRXNexus: nexus nexusList

 $May\ contain\ Character\ data\ only$

Content model <content> <textNode/></content>

Schema Declaration element note text

> This element marks the beginning of a new page or page-like section on the text-bearing surface.

Module derived-module-KRX

Attributes

@ed Identifier of the edition to which this page belongs. Status Optional
Datatype string
@n Number or other label used to refer to this page. Status Optional
Datatype string
@xml:id Status Recommended
Datatype ID

KRXToken: tg

May contain Empty element Content model <content> <empty/></content> Schema Declaration

```
element pb
{
    attribute ed { text }?,
    attribute n { text }?,
    attribute xml:id { text }?,
    empty
}
```

<resp> Person responsible for some aspect of the work.

Module KRXManifest

Attributes

@role Status Optional

 $Datatype \ string$

Sample values include: author Author

compiler Compiler

translator Translator

©key A key identifying this person in some reference system. Status Optional

Datatype string

Contained by creation

May contain Character data only

Content model <content> <textNode/></content>

Schema Declaration

element resp { attribute role { text }?, attribute key { text }?, text }

 $\langle t \rangle$ A token. Module KRXToken Attributes **@role** Token type. Status Required Legal values are: h Token is part of a heading. **p** Token is part of a paragraph. **s** Token is part of a seg element. **n** Token is part of a note or annotation of any kind. **q** Token is part of a quotation. **v** Token is part of a verse line. o Token is part of a textual feature not in this list. **Opos** The sequencial number of this token within this element (or token type). Status Optional Datatype nonNegativeInteger **Qtp** The sequencial number of this token within the whole text. Status Required Datatype nonNegativeInteger @f Punctuation or other non-token text items, immediately following the token. Status Optional Datatype string **Op** Punctuation or other non-token text items, immediately preceding the token. Status Optional Datatype string On Label or identifier of the element in the text of which this token is part. If none is available, the code generating the token file should make one up on the fly. Status Required Datatype string **@cp** Codepoint of the token character. Status Optional Datatype nonNegativeInteger **Oposition** position and content of marks out of line, but related to this token. The description is similar to CSS description on HTML @style: 'left: か;' would indicate a p syllable to the left of this token. Status Optional Datatype string @kundokuten Kundoku marks related to this token. Status Optional Datatype string **@ruby** Pronounciation marks related to this token. Status Optional

Datatype string

KRXToken: tg May contain Character data only Content model <content> <textNode/></content> Schema Declaration

```
element t
{
    attribute role { "h" | "p" | "s" | "n" | "q" | "v" | "o" },
    attribute pos { text }?,
    attribute tp { text },
    attribute f { text }?,
    attribute p { text }?,
    attribute n { text },
    attribute cp { text }?,
    attribute position { text }?,
    attribute kundokuten { text }?,
    attribute ruby { text }?,
    text
}
```

 ${<}tList{>}$ Root for token that may contain one or more ${<}tg{>}$ elements.

Module KRXToken Attributes ©xml:id Status Recommended Datatype ID ©ed Reference to the edition defined in the manifest. Status Required Datatype string ©n A label

Status Optional

Datatype string

©fileseq If the tokens are in several files, this gives the sequential number of the file.

Status Optional

 $Datatype \ {\rm nonNegativeInteger}$

May contain KRXToken: tg Content model

<content>
 <elementRef key="tg" maxOccurs="unbounded"/>
</content>

 $Schema\ Declaration$

```
element tList
{
    attribute xml:id { text }?,
    attribute ed { text },
```

```
attribute n { text }?,
attribute fileseq { text }?,
krx_tg+
}
```

<tg> A group of tokens. Module KRXToken Attributes **@xml:id** The identifier of this token group. Status Optional Datatype ID **@n** A label. Status Optional Datatype string **Orole** Token group type. Status Optional Legal values are: h Token group is a heading. **p** Token group is (part of) a paragraph. **s** Token group is a seg element. **n** Token group is (part of) a note or annotation of any kind. **q** Token group is (part of) a quotation. **v** Token group is (part of) a verse line. **o** Token group is (part of) a textual feature not in this list. Oposition position and content of marks out of line, but related to this token. The description is similar to CSS description on HTML @style: 'left: か;' would indicate a m syllable to the left of this token. Status Optional Datatype string @kundokuten Kundoku marks related to this token. Status Optional Datatype string **@ruby** Pronounciation marks related to this token. Status Optional Datatype string *KRXToken:* tList tg May contain *KRXToken:* t tg *derived-module-KRX:* lb pb Content model <content> <alternate maxOccurs="unbounded"</pre> minOccurs="0">

```
maxOccurs="unbounded" minOccurs="0"/>
  <elementRef key="lb"
    maxOccurs="unbounded" minOccurs="0"/>
  </alternate>
  </content>
```

Schema Declaration

```
element tg
{
    attribute xml:id { text }?,
    attribute n { text }?,
    attribute role { "h" | "p" | "s" | "n" | "q" | "v" | "o" }?,
    attribute position { text }?,
    attribute kundokuten { text }?,
    attribute ruby { text }?,
    ( krx_tg* | krx_t* | krx_pb* | krx_lb* )*
}
```

<title> Title of the work.

Module KRXManifest

 $Contained \ by$ description edition Group manifest

May contain Character data only

Content model <content> <textNode/></content>

Schema Declaration element title text

< tokenmap > Mappings from textual features to token types.

Module KRXManifest Contained by edition May contain derived-module-KRX: map Content model

```
<content>
<elementRef key="map"
maxOccurs="unbounded" minOccurs="1"/>
</content>
```

Schema Declaration | element tokenmap krx_map+