

Veri Temizliği ve DSpace'e Toplu Kayıt Aktarımı

Anahtar Kelimeler: Access, LibreOffice, CSV, OpenRefine,
dSpace, SimpeArchiveFormat, SAFbuilder

Zeki Çelikbaş
İTÜ, celikbas@itu.edu.tr

4. ULUSAL AÇIK ERİŞİM ÇALIŞTAYI

19-21 EKİM 2015, TÜBİTAK, Ankara

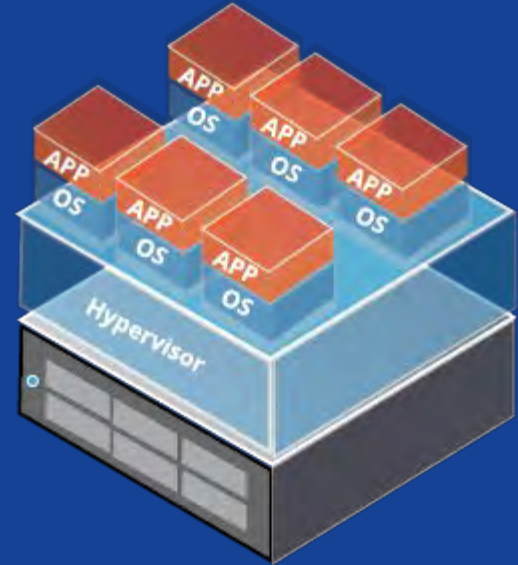
National Open Access Symposium 19-21 October 2015

dSpace Kurulumu

İTÜ Kütüphane



İTÜ BİDB

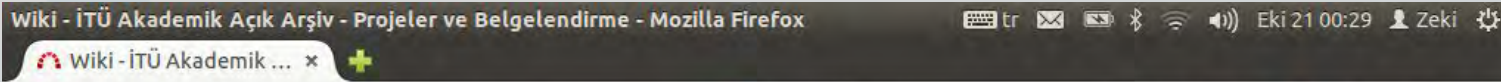


Polen Sunucu 1

Test Sunucu 2

Test Sunucu 3

Belgeleme Sunucusu



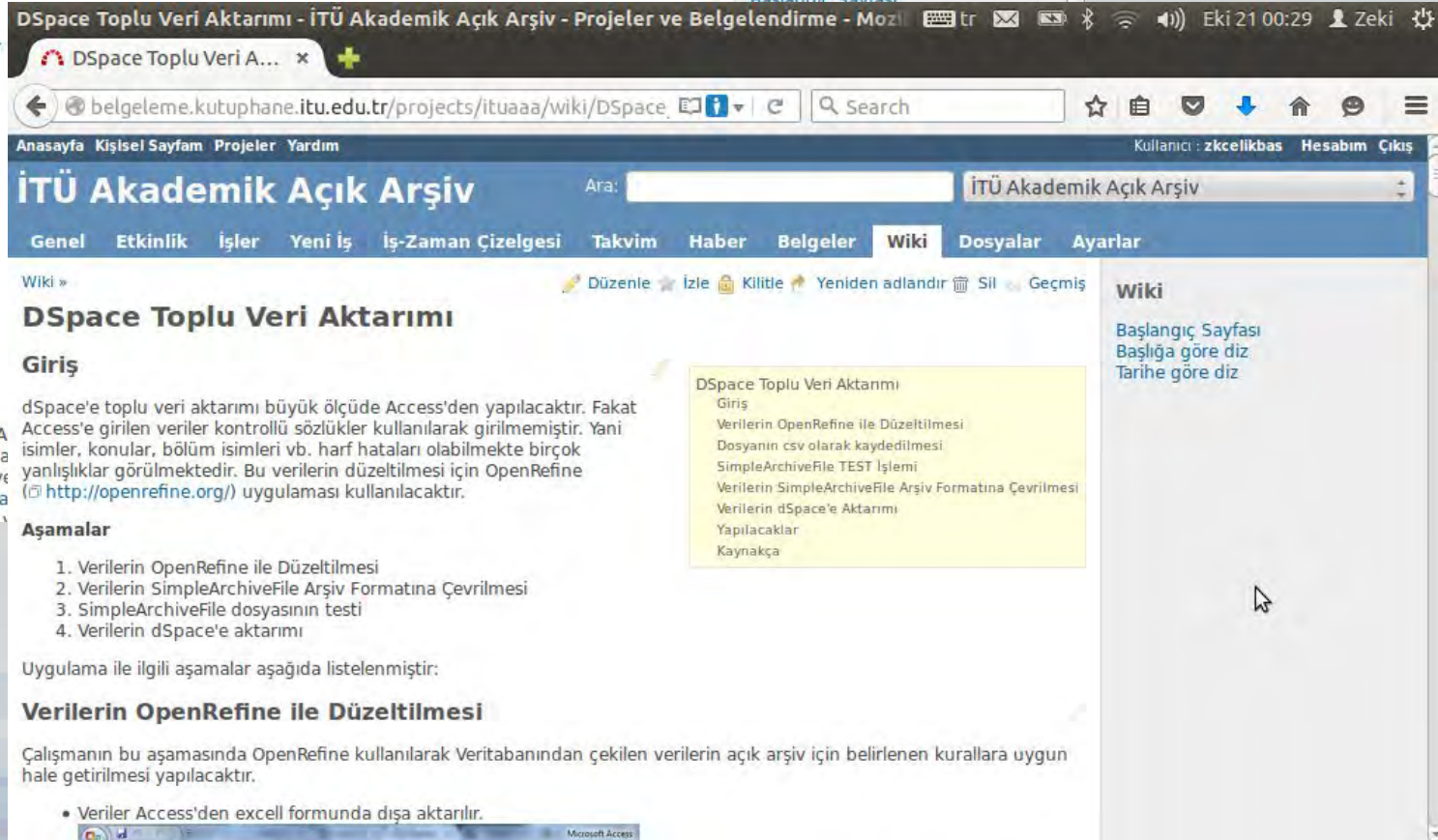
- ### Çalışmalar
- dSpace'e Sonradan Eklenen Alanlar
 - dSpace Toplu Veri Aktarımı
 - dSpace Toplu Veri Aktarımı Güncel
 - Kullanılan Uygulamalar
 - Lisans'lar
 - Tasarım önerileri
 - Polen'e Gönderilen Kayıtlar
 - Çeviri düzeltmeleri
 - Sunucu ile ilgili işlemler

Toplantı Notları

- Notlar - 07.11.2014
- Toplantı Notları - 13.04.2015
- Toplantı Notları - 11.05.2015

Yapılacaklar

- MarcEdit ile dSpace sunucusuna OAI request to that OAI-PMH server /oai/request?verb=ListMetadata the DSpace Demo Server by



DSpace Toplu Veri Aktarımı

Giriş

dSpace'e toplu veri aktarımı büyük ölçüde Access'den yapılacaktır. Fakat Access'e girilen veriler kontrollü sözlükler kullanılarak girilmemiştir. Yani isimler, konular, bölüm isimleri vb. harf hataları olabilmekte birçok yanlışlıklar görülmektedir. Bu verilerin düzeltilmesi için OpenRefine (<http://openrefine.org/>) uygulaması kullanılacaktır.

Aşamalar

1. Verilerin OpenRefine ile Düzeltilmesi
2. Verilerin SimpleArchiveFile Arşiv Formatına Çevrilmesi
3. SimpleArchiveFile dosyasının testi
4. Verilerin dSpace'e aktarımı

Uygulama ile ilgili aşamalar aşağıda listelenmiştir:

Verilerin OpenRefine ile Düzeltilmesi

Çalışmanın bu aşamasında OpenRefine kullanarak Veritabanından çekilen verilerin açık arşiv için belirlenen kurallara uygun hale getirilmesi yapılacaktır.

- Veriler Access'den excell formunda dışa aktarılır.

DSpace Toplu Veri Aktarımı

Giriş

- Verilerin OpenRefine ile Düzeltilmesi
- Dosyanın csv olarak kaydedilmesi
- SimpleArchiveFile TEST işlemi
- Verilerin SimpleArchiveFile Arşiv Formatına Çevrilmesi
- Verilerin dSpace'e Aktarımı
- Yapılacaklar
- Kaynakça

Kısa Tarihçe

- ◆ 2002 FBE Tez veri tabanı
- ◆ 2007-2013 Online Tez Veri Tabanı
- ◆ 2013-2015 Yök Tez Veri Tabanı
- ◆ 2015 - İTÜ Polen



~12.000 Kayıt

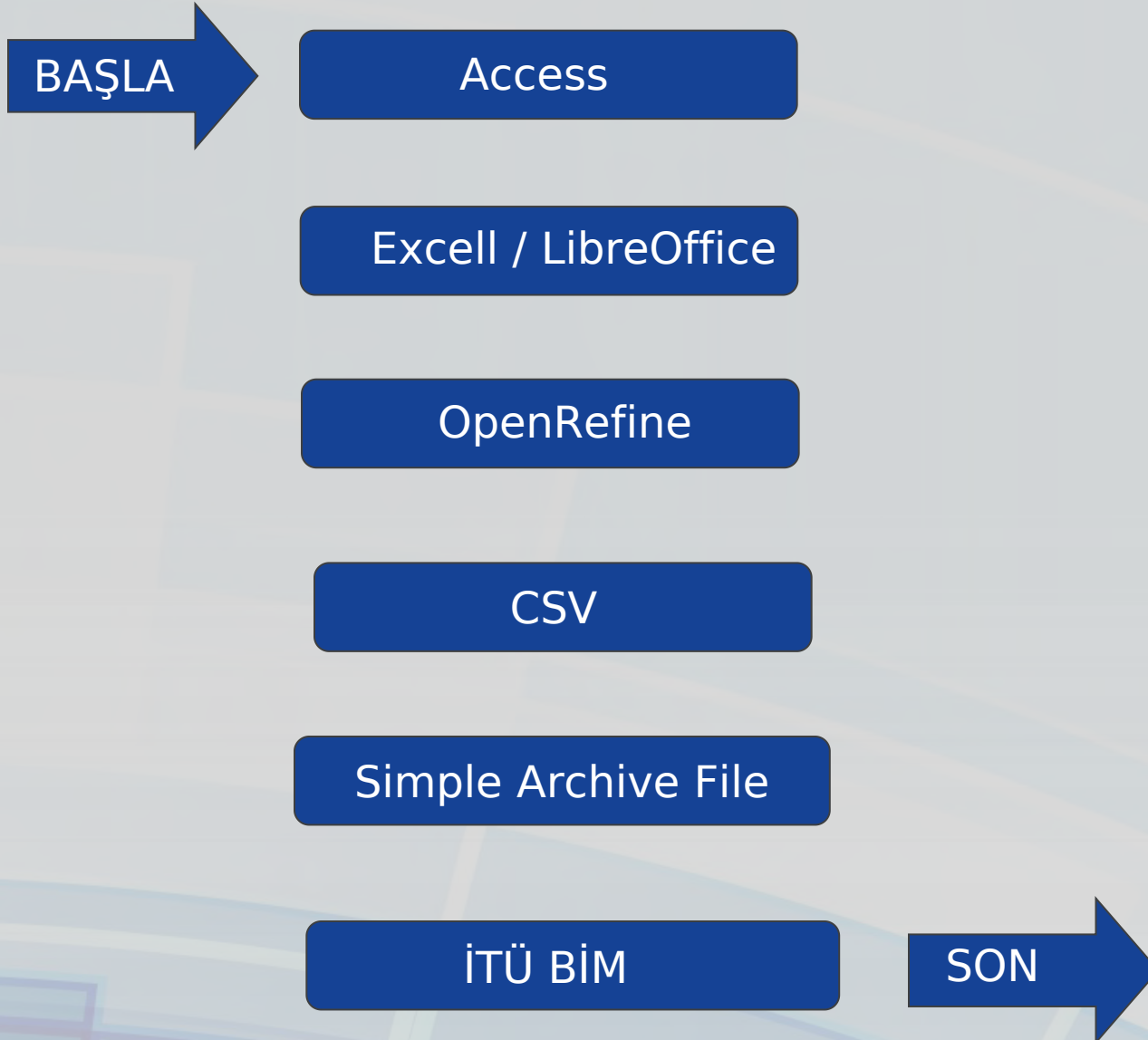
İTÜ Kütüphanee



~9.000 Kayıt

$$\frac{9000}{(10 - 4) \times 22} = 115$$

İş Akışı Özeti



Sistemin Kurulumu

Kullanılan Uygulamalar



- **JDK** kurulumu: * JDK kurulumu yaptıktan sonra JAVA_HOME değişkeninin sisteme eklenmesi gerekmektedir. Aşağıda maven kurulumu sırasında bu işlemi de yapınız.

- **Apache Maven** (maven)

<https://maven.apache.org/>

- **Cygwin** bilgisayarda linux ortamı yaratmak için kurulur.

<http://cygwin.com/install.html>

- **SAFBuilder**: "Simple Archive Format" dosyası oluşturmak için kurulur. SAFBuilder, kurulum gerektirmez herhangi bir dizine açılması yeterlidir. Fakat çalışması için JDK ve Maven'e ihtiyaç duyar.

<https://github.com/peterdietz/SAFBuilder>

- **OpenRefine** (eski googlerefine)

<http://openrefine.org/>



Yardımcı Programlar

PDFsam: Pdf düzenleme (kesme, yapıştırma, bölme, vb.)

<http://www.pdfsam.org/>



Access ==> LibreOffice (Excell)

enstitu_t ==> dc.publisher[tr_TR]
enstitu_i ==> dc.publisher[en_US]
program_t ==> dc.contributor.department[tr_TR]
program_i ==> dc.contributor.department[en_US]
tezdanisman ==> dc.contributor.advisor
tezturu_t ==> dc.description.degree[tr_TR]
tezturu_i ==> dc.description.degree[en_US]
yil_t ==> dc.date
tezciadi ==> dc.contributor.author
baslik_t ==> dc.title[tr_TR]
baslik_i ==> dc.title.alternative[en_US]
abstract_t ==> dc.description.abstract[tr_TR]
abstract_i ==> dc.description.abstract[en_US]
AnahKel ==> dc.subject[tr_TR]
Keyword ==> dc.subject[en_US]
today ==> dc.date.accessioned
onaytarihi ==> dc.date.issued
yokref_id ==> dc.contributor.authorID
[YENİ SÜTUN] ==> dc.type[en_US]
[YENİ SÜTUN] ==> dc.type[tr_TR]
[YENİ SÜTUN] ==> dc.description[tr_TR]
[YENİ SÜTUN] ==> dc.description[en_US]
[YENİ SÜTUN] ==> dc.rights[tr_TR]
[YENİ SÜTUN] ==> dc.rights[en_US]

Eklenen Sabit Alanlar

dc.contributor.authorID = yokref_id

dc.type[en_US] = Thesis
dc.type[tr_TR] = Tez

dc.description[tr_TR] = Tez (Yüksek Lisans) -- İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, [DATE]
dc.description[en_US] = Thesis (M.Sc.) -- İstanbul Technical University, Institute of Science and Technology, [DATE]

dc.rights[tr_TR] = İTÜ tezleri telif hakkı ile korunmaktadır. Bunlar, bu kaynak üzerinden herhangi bir amaçla görüntülenebilir, ancak yazılı izin alınmadan herhangi bir biçimde yeniden oluşturulması veya dağıtılması yasaklanmıştır.

dc.rights[en_US] = İTÜ theses are protected by copyright. They may be viewed from this source for any purpose, but reproduction or distribution in any format is prohibited without written permission.



*A free, open source, powerful
tool for working with messy data*

<http://openrefine.org/>

Essential: These are the features you'd use 80% of the time when you use Refine. They help you clean up your data, extend it, and export it out for other tools to consume.

- Importing

- Filtering / faceting

- Editing:

 - Editing cells, editing cells by Clustering

 - Editing columns, creating columns by Extending data

 - Editing rows

 - Understanding expressions

 - Understanding regular expressions

- Exporting

- History (undo/redo)

OpenRefine Veri Düzenleme

Mekatronik_Yukseklisans2.xlsx - Google Refine - Iceweasel

Mekatronik_Yukseklisa... x

127.0.0.1:3333/project?project=2439197318494

Google refine Mekatronik_Yukseklisans2.xlsx Permalink

Open... Export Help

Facet / Filter Undo / Redo 2

90 rows Extensions: Freebase

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 10 next > last »

▼ All ▼ dc_publisher(tr_ ▼ dc_publisher(en ▼ program_t ▼ dc_contributor_ ▼ tezturu_t ▼ tezturu_i ▼

1.	Fen Bilimleri Enstitüsü	Institute of Science and Technology	Mekatronik	Doç. Dr. Metin GÖKAŞAN	Yüksek Lisans	M.Sc.
----	-------------------------	-------------------------------------	------------	------------------------	---------------	-------

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

OpenRefine Veri Düzenleme

Facet

Text filter

Edit cells

Edit column

Transpose

Sort...

View

Reconcile

Transform...

Common transforms

Fill down

Blank down

Split multi-valued cells...

Join multi-valued cells...

Cluster and edit

Cluster & Edit column "dc_contributor_advisor"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method nearest neighbor

Distance Function levenshtein

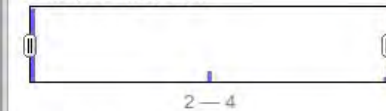
Radius 1.0

Block Chars 6

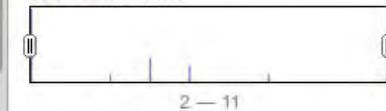
17 clusters found

Cluster Size	Row Count	Values In Cluster	Merge?	New Cell Value
4	6	- Doç. Dr. Hakan TEMELTAŞ (3 rows) - Doç.Dr. Hakan Temeltaş (1 rows) - Doç. Dr. HAKAN TEMELTAŞ (1 rows) - Doç. Dr. Hakan Temeltaş (1 rows)	☒	Doç. Dr. Hakan TEMELTAŞ
3	5	- Doç. Dr. Hakan TEMELTAŞ (3 rows) - Doç.Dr. Hakan Temeltaş (1 rows) - Doç. Dr. Hakan Temeltaş (1 rows)	☒	Doç. Dr. Hakan TEMELTAŞ
3	8	- Yrd. Doç. Dr. Levent OVACIK (4 rows) - Yrd. Doç. Dr. Levent Ovacik (3 rows) - Yrd.Doç. Dr. Levent OVACIK (1 rows)	☒	Yrd. Doç. Dr. Levent OVACIK
2	11	- Prof. Dr. Levent Güvenç (8 rows) - Prof. Dr. Levent GÜVENÇ (3 rows)	☒	Prof. Dr. Levent Güvenç
2	2	- Doç.Dr. Bilin Aksun Güvenç (1 rows) - Doç. Dr. Bilin Aksun Güvenç (1 rows)	☒	Doç.Dr. Bilin Aksun Güvenç
2	2	- Prof. Dr. Metin Gökaşan (1 rows) - Prof. Dr. Metin GÖKAŞAN (1 rows)	☒	Prof. Dr. Metin Gökaşan

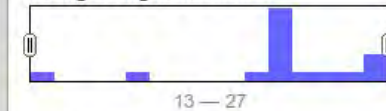
Choices in Cluster



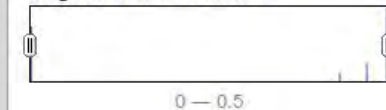
Rows in Cluster



Average Length of Choices



Length Variance of Choices



Select All

Deselect All

Merge Selected & Re-Cluster

Merge Selected & Close

Close



OpenRefine Veri Düzenleme

Custom text transform on column dc_contributor_advisor

Expression Language Google Refine Expression Language (GREL) ▼

```
value.replace('Yrd.', '').replace('Yard.', '').replace('Y.D', 'D').replace('Doç.', '').replace('Dr.', '').replace('Prof.', '').replace('Asst.', '').replace('Öğr. Gör.', '').trim()
```

No syntax error.

Preview History Starred Help

1.	Doç. Dr. Metin GÖKAŞAN	Metin GÖKAŞAN
2.	Prof. Dr. Levent Güvenç	Levent Güvenç
3.	Yrd. Doç. Dr. M. Turan Söylemez	M. Turan Söylemez

On error ☒ keep original ☐ set to blank ☐ store error ☐ Re-transform up to times until no change

OK Cancel



OpenRefine Veri Düzenleme

Custom text transform on column dc_contributor_advisor

Expression

Language **jython**

```
words = value.split()
last = words.pop()
fname = ' '.join(words)
return last + ", " + fname
```

No syntax error.

Preview

History

Starred

Help

row	value	words = value.split() last = words.pop() fname = ' '.join(words) return last + ", " + fname
1.	Metin GÖKAŞAN	GÖKAŞAN, Metin
2.	Levent Güvenç	Güvenç, Levent
3.	M. Turan Söylemez	Söylemez, M. Turan
4.	Hakan TEMELTAŞ	TEMELTAŞ, Hakan
5.	Levent OVACIK	OVACIK, Levent

On error

- ☒ keep original
☐ set to blank
☐ store error

☐ Re-transform up to times until no change

OK

Cancel



OpenRefine Veri Düzenleme

dc_contributor_ testuru_t testuru_i dc

Facet

Text filter

Edit cells ▶ Transform...

Edit column ▶ Common transform

Transpose ▶ Fill down

Sort... Blank down

View ▶ Split multi-valued c

Reconcile ▶ Join multi-valued c

Cluster and edit...

Custom text transform on column dc_contributor_advisor

Expression

Language jython

```
result = ""
for word in value.split():
    if word.startswith("İ"):
        word = "i" + word[1:]
    word = word.capitalize()
    result = result + " " + word
return result
```

No syntax error.

Preview

History

Starred

Help

- | | | |
|----|--------------------|--------------------|
| 1. | GÖKAŞAN, Metin | Gökaşan, Metin |
| 2. | Güvenç, Levent | Güvenç, Levent |
| 3. | Söylemez, M. Turan | Söylemez, M. Turan |
| 4. | TEMELTAŞ, Hakan | Temeltaş, Hakan |
| 5. | OVACIK, Levent | Ovacik, Levent |
| 6. | KIITILİ Kenan | Kutlu Kenan |

On error

- ☒ keep original
- ☐ set to blank
- ☐ store error

☐ Re-transform up to 10 times until no change

OK

Cancel



OpenRefine Veri Düzenleme

- Büyük Yazılan Başlıkları
OK > Edit Cell > Common Transforms > To titlecase
- dc.title.alternative[en_US]
OK > Edit Cell > Transform
value.replace('ı','i')
- Başlık, Konu başlıklarında ve Özetlerde (Türkçe ve İngilizce)
OK > Edit Cell > Common Transforms > Collapse consecutive whitespace
dc.title.alternative[en_US]: value.replace(/\s+/, ' ')
- dc.subject[tr_TR] ve dc.subject[en_US]
OK > Edit Cell > Transform...
Language: Google Refine Expression Language (GREL)
value.replace(' ', '|||').replace(' ', '|||')



Veri Kontrol

Google Refine'de düzenlemebittikten sonra dosyanın excel formatında dışarı aktarılması gerekir. Bu işlem için

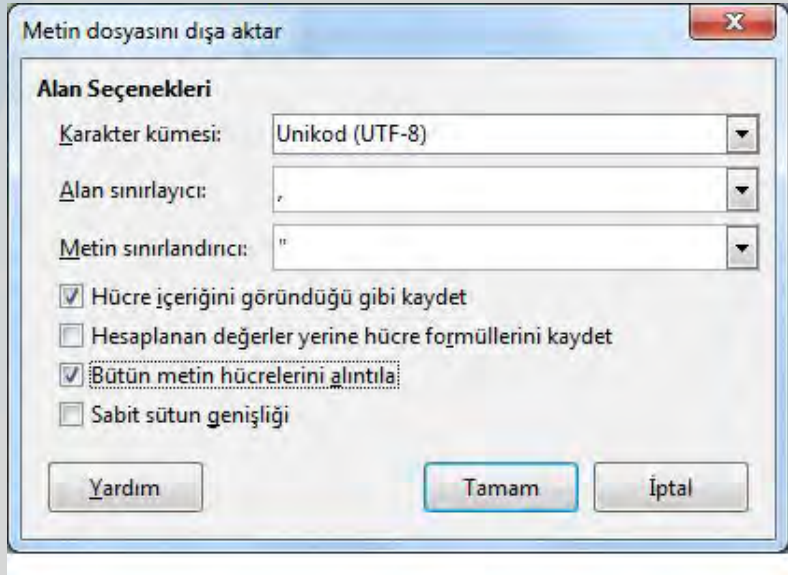
"Export > Excel" seçilir.

Exel dosyasında son kontroller yapılır:

- Danışman adı eş danışman adıyla verilebilir.Bu alana sadece danışman adını giriyoruz. Kütüphane kataloğundan kontrol edilerek düzeltilebilir.
- Yazar(Danışman ve Yazar) adlarının önünde fazla boşluklar olabilir,silinmesi gerekiyor.
- Özetlerde (Türkçe ve İngilizce) özel karakterler bozuk olabilir, düzeltilecek.



LibreOffice CSV Dışarı Aktar



Kayıt türü ==> Metin CSV (.csv)(*.csv)

Karakter kümesi: Unicode (UTF-8)

Alan sınırlayıcı: ,

Metin sınırlayıcı: "

Hücre içeriğini görüldüğü gibi kaydet: SEÇİLİ

Bütün metin hücrelerini alıntıla: SEÇİLİ

Simple Archive Format Packager

<https://wiki.duraspace.org/display/DSPACE/Simple+Archive+Format+Packager>

<https://github.com/peterdietz/SAFBuilder>

◆ Giriş

1. CSV file
2. pdf, jpg, txt documanlar
3. lisans dökümanı

◆ Çıkış

1. XML Biçimlendirilmiş veriler
2. Dizin
3. Zip dosya

Verilerin SimpleArchiveFile Arşiv Formatına Çevrilmesi

```
zeki@kitaplik3:~/bin/SAFBuilder$ ./safbuilder.sh -c  
derleme/Mekatronik_YuksekLisans.csv
```

```
Recompiling SAFBuilder, just a moment...  
Detected input CSV as:UTF-8  
Opened CSV File:derleme/Mekatronik_YuksekLisans.csv  
Output directory is:  
/home/zeki/bin/SAFBuilder/derleme/SimpleArchiveFormat  
File: Mekatronik_YuksekLisans.csv has been used 0 times.
```

```
zeki@kitaplik3:~/bin/SAFBuilder$ ls  
Mekatronik_YuksekLisans.zip
```

dublin_core.xml

dublin_core.xml x

```
<?xml version="1.0" encoding="UTF-8"?>
<dublin_core>
  <dcvalue element="publisher" language="tr_TR">Fen Bilimleri Enstitüsü</dcvalue>
  <dcvalue element="publisher" language="en_EN">Institute of Science and Technology</dcvalue>
  <dcvalue element="contributor" qualifier="advisor">Güvenç, Levent</dcvalue>
  <dcvalue element="type" language="tr_TR">Yüksek Lisans</dcvalue>
  <dcvalue element="type" language="en_EN">M.Sc.</dcvalue>
  <dcvalue element="date">2011</dcvalue>
  <dcvalue element="contributor" qualifier="author">Karaahmetoğlu, Raif</dcvalue>
  <dcvalue element="title" language="tr_TR">Gcdc 2011 Yarı Otonom Kooperatif Adaptif Sürüş Yarış Aracının Tasarımı</dcvalue>
  <dcvalue element="title" qualifier="alternative" language="en_EN">Design Of Gcdc 2011 Semi-autonomous Cooperative Adaptive Driving
  <dcvalue element="description" qualifier="abstract" language="tr_TR">Bu çalışmada, Büyük Kooperatif Sürüş Yarışması için yarı otonom
  <dcvalue element="description" qualifier="abstract" language="en_EN">In this study, semi-autonomous vehicle was generated for Grand
  <dcvalue element="subject" language="tr_TR">cacc</dcvalue>
  <dcvalue element="subject" language="tr_TR">kooperatif adaptif seyir kontrolcüsü</dcvalue>
  <dcvalue element="subject" language="tr_TR">büyük kooperatif sürüş yarışması</dcvalue>
  <dcvalue element="subject" language="tr_TR">otonom araçlar</dcvalue>
  <dcvalue element="subject" language="en_EN">cacc</dcvalue>
  <dcvalue element="subject" language="en_EN">cooperative adaptive cruise control</dcvalue>
  <dcvalue element="subject" language="en_EN">grand cooperative driving challenge</dcvalue>
  <dcvalue element="subject" language="en_EN">autonomous vehicles</dcvalue>
  <dcvalue element="contributor" qualifier="authorID">410126</dcvalue>
</dublin_core>
```


	A	B	C	D	E
1	filename	dc.publisher[tr_TR]	dc.publisher[en_EN]	program t	dc.contributor.advisor
2	15000001.pdf_bundle:ORIGINAL license.txt_bundle:LICENSE	Fen Bilimleri Enstitüsü	Institute of Science and Technology	Mekatronik	Güvenç, Levent
3	15000002.pdf_bundle:ORIGINAL license.txt_bundle:LICENSE	Fen Bilimleri Enstitüsü	Institute of Science and Technology	Mekatronik	Ovacık, Levent
4	15000003.pdf_bundle:ORIGINAL license.txt_bundle:LICENSE	Fen Bilimleri Enstitüsü	Institute of Science and Technology	Mekatronik	Kurtulan, Salman
5	15000004.pdf_bundle:ORIGINAL license.txt_bundle:LICENSE	Fen Bilimleri Enstitüsü	Institute of Science and Technology	Mekatronik	Özkol, Ibrahim
6	15000005.pdf_bundle:ORIGINAL license.txt_bundle:LICENSE	Fen Bilimleri Enstitüsü	Institute of Science and Technology	Mekatronik	Güvenç, Levent
7	15000006.pdf_bundle:ORIGINAL license.txt_bundle:LICENSE	Fen Bilimleri Enstitüsü	Institute of Science and Technology	Mekatronik	Güvenç, Levent
8	15000007.pdf_bundle:ORIGINAL license.txt_bundle:LICENSE	Fen Bilimleri Enstitüsü	Institute of Science and Technology	Mekatronik	Sönmez, Ümit
9	15000008.pdf_bundle:ORIGINAL license.txt_bundle:LICENSE	Fen Bilimleri Enstitüsü	Institute of Science and Technology	Mekatronik	Öke, Gülav
10					

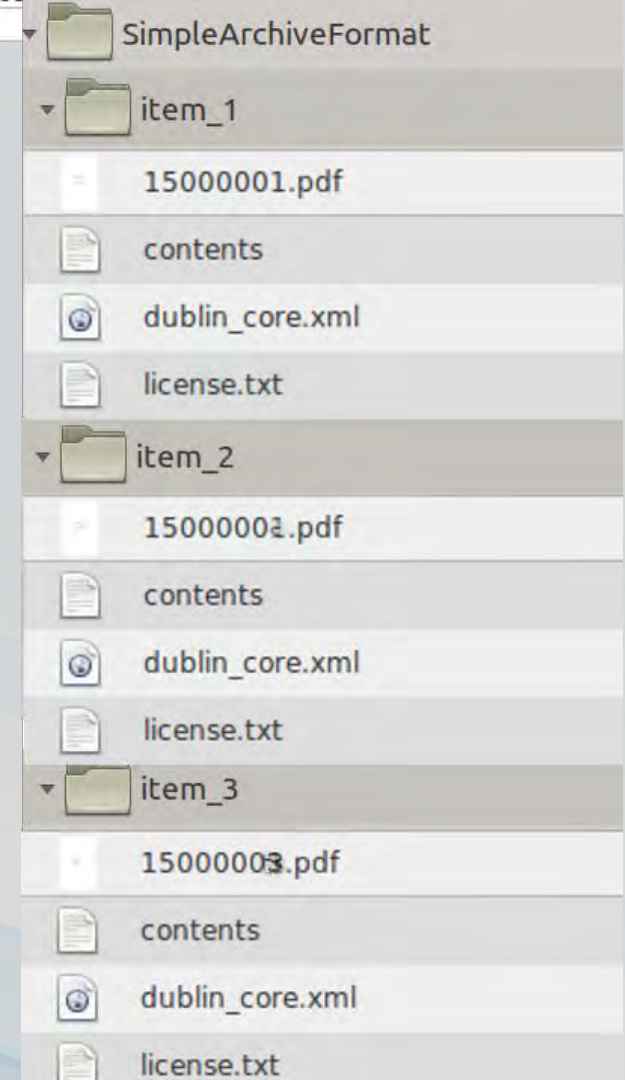
contents

15000001.pdf bundle:ORIGINAL
license.txt bundle:LICENSE

license.txt

DAGITIM LISANSI

Bu lisansi imzalayip göndererek, gönderinizi dünya capında basılı halde, elektronik ortamda ve herhangi başka bir kanalla yeniden oluşturmak, çevirisini yapmak (asagıda tanımlandığı üzere) ve/veya dağıtmak için İstanbul Teknik Üniversitesi'nin (ITU) özel olmayan haklarını kabul etmiş olursunuz.....



dSpace İçeri Aktar (import)

```
sudo /dspace/bin/dspace import -a  
-e polen@itu.edu.tr  
-c 11527/110  
-s /home/dspace/SAFBuilder/import/SimpleArchiveFormat/  
-m /home/dspace/SAFBuilder/imported/11527-110-ek1.map
```

From: library@itu
To: bim@itu

Ek olarak yüklenecek.
İyi çalışmalar.
<https://polen.itu.edu.tr/handle/11527/110>
http://web.itu.edu.tr/zsolta/dspace/Metalurji_ve_Malzeme_Muhendisligi_DR.zip

From: bim@itu
To: library@itu

> Merhaba,
>
> Ekledim.
>
> İyi çalışmalar.

Sonuçlar ve Yapılacaklar

SONUÇLAR

- ◆ Çok miktarda veri kısa zamanda sisteme girilmiştir.
- ◆ FBE verilerinin girişi bitmek üzeredir.
- ◆ Diğer enstitülerle benzer çalışmalar planlanmakta ve bu doğrultuda Sosyal Bilimler Enstitüsü verileri kütüphaneye ulaşmıştır.

YAPILACAKLAR

- ◆ Enstitülerden otomatik veri çekmek için servis yazılımları üzerine çalışılacak.
- ◆ SWORD protokolü ile bağlantı çalışmaları yapılacak
- ◆ Petek sisteminde bulunan kayıtlar dspace'e aktarılacak.

Teşekkürler...

