Klaus A J Riederer, Dr. (tech),
KAR Ltd, CEO:

# Seeing is believing: how to make OCR work in practice

KAR

WE HAVE JUST MOVED – WE ARE A FACTORY IN:
Suurpellontie 1, FI-07700 Koskenkylän saha, Finland
www.kar.fi
kar@kar.fi

KAR Oy Ab Ltd GmbH S.A.R.L
Valokuvavalmistamo KAR, Photography Production KAR, Bildtillverkning KAR, FotografieFertigung KAR, Photographie production KAR
Taidevedostamo KAR, Fine Art Printmaking KAR, Konsttryckeri KAR, Kunstdruckwerk KAR, Imprimerie des Beaux-Arts KAR
KAR Audio

KAR

# -1. Lets make a poll!

- Please answer truthfully by raising your hands:
  - 1. Who believes in **free lunches**?
  - 2. Who believes in **miracles**?
  - 3. Who believes everything must be **digitized**?
  - 4. Who believes everything can be turned into **bits** (and pieces)?
  - 5. Who has seen a proper **OCR (optical character recognition)** done in an automated fashion?

# o. take the FREE LUNCH!

KAR

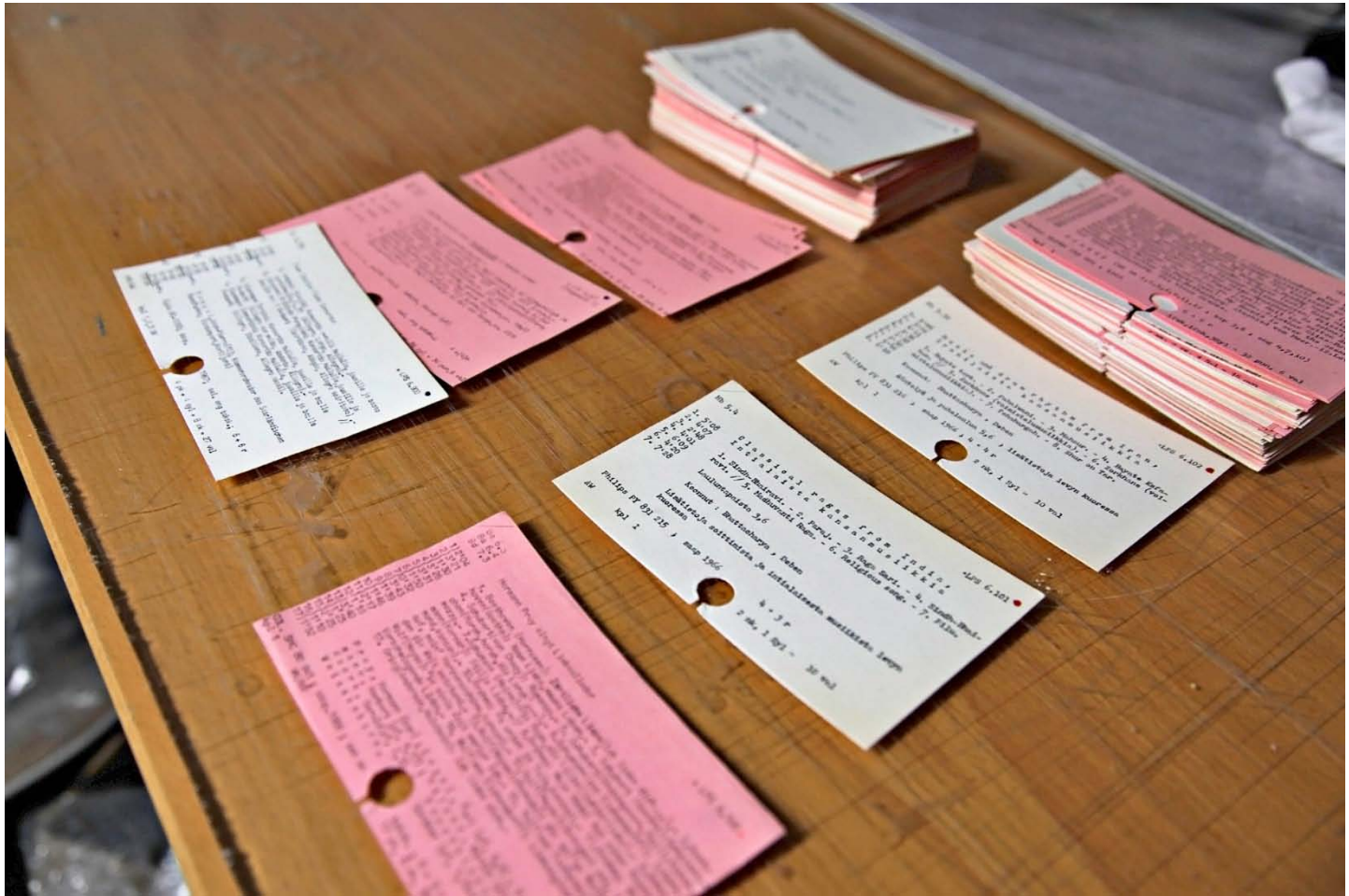
- If somebody wants to offer you a free lunch, take it!
  - 2012 BAAC conference excursion to the Finnish Broadcasting Company (Yleisradio, **YLE**)

Library cabinets full of cardboards – YLE vinyl record archives

# 0+1. I said: take the FREE LUNCH!

KAR

- If somebody wants to offer you a free lunch, take it!
  - We have 37 000 cardboards from our vinyl archives, ca. 1966 – 1973. Care to scan & **OCR** them, as nobody else cannot do it?

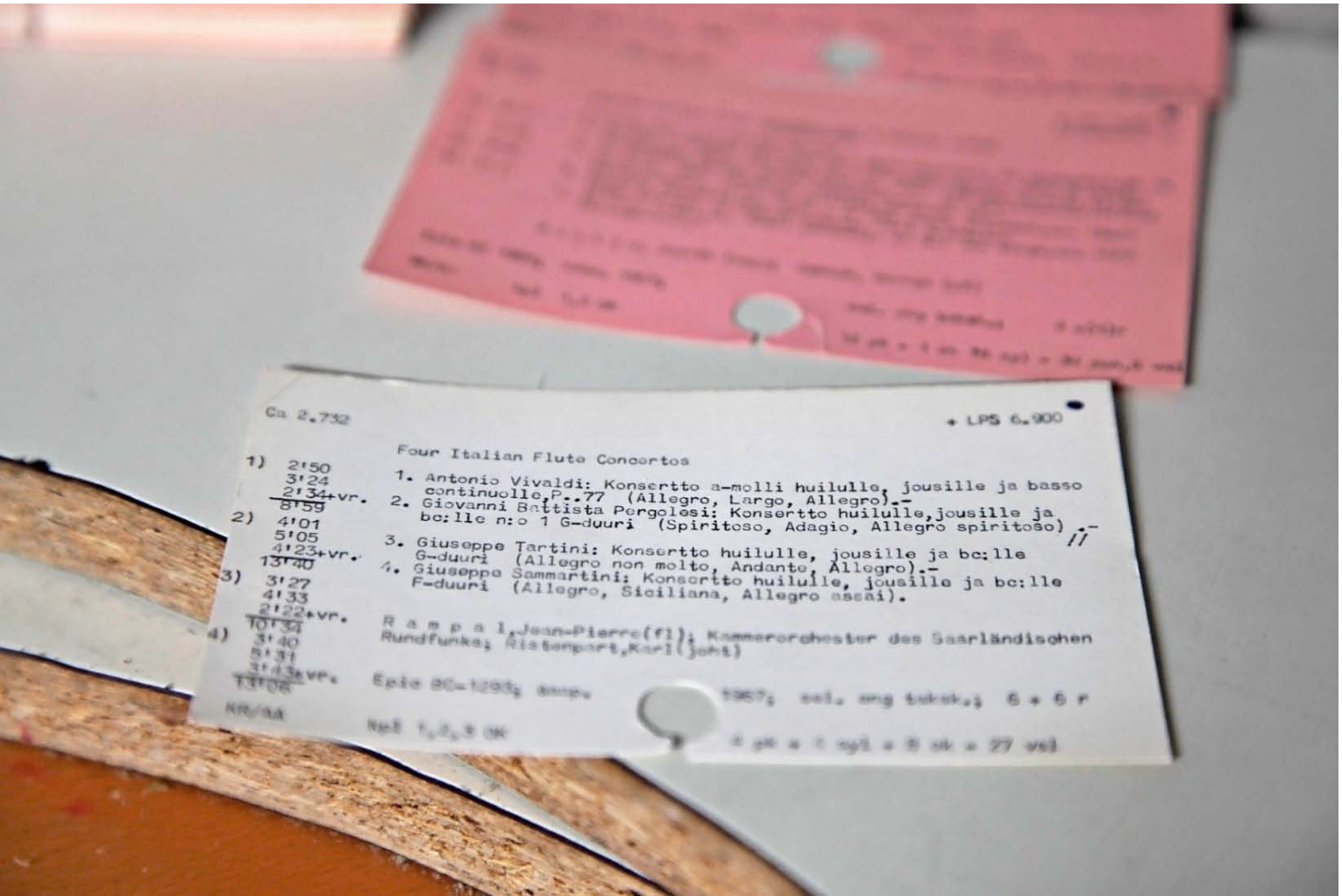


They are, of course, all different (as such a computer would regard them)!

KAR

# 1. What do you want?

- What the customer wants?
  - Customer's initial specification

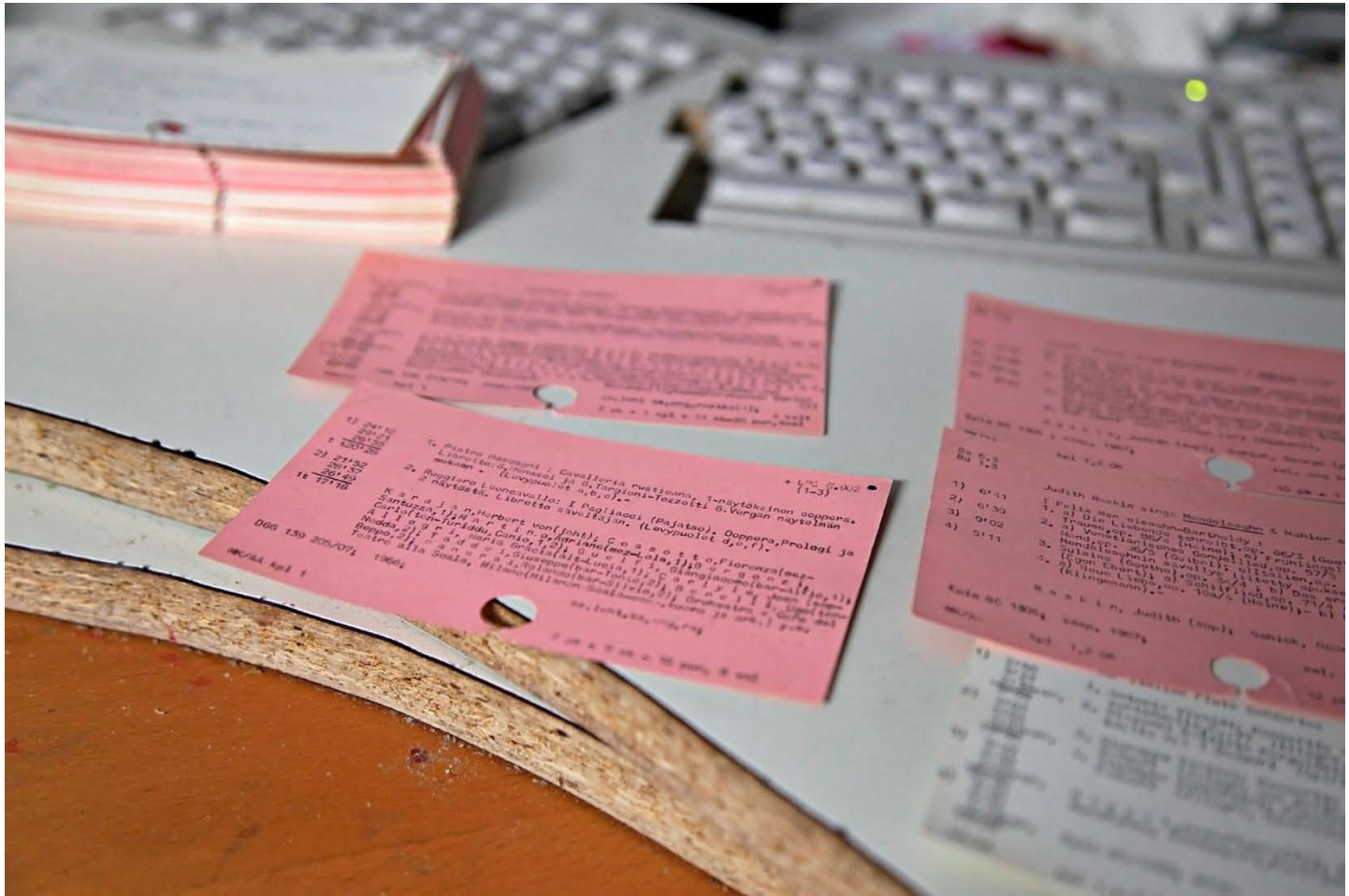


1. Take all the data out of the printed card
2. Turn the data into information in XML format (for YLE's METRO database system) **= structured information**

KAR

# 2. What do you need?

- What the customer wants and (definitely) needs are not the same thing!
  - Do you really need all the data (even the obsolete, e.g. archive writer initials)



**What are your minimum demands?**

# 3.YLE OCR case: make a pilot!

KAR

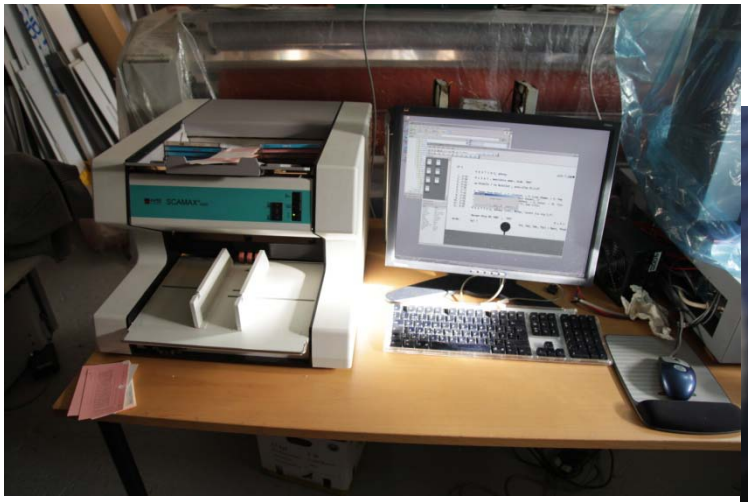
- The entrepreneur takes all the risks (bah!), so take the first 100 cards, scan & OCR them as a pilot!
  - But first, clean and tweak your machines

*My benefit is your benefit and your my benefit!*

KAR

# 4.YLE OCR case: make a ~~test~~ pilot!

- Testing one-two-three;
  alfa, beta, gamma;
  o-series – when does this "pilot" actually start?



- Scanning in 400 dpi (1:1),
  black-and-white uncompressed TIFs

# 4.YLE OCR case: scanning

KAR

- Production scanning – show video or photos

KAR

# 5. Optical character recognition (OCR)

- What is OCR?
- How does it work?
- How well does it work?
- How much does it cost?
- OCR – many readymade (and less ready) solutions
  - Tesseract (open source, from HP originally)
  - *Abbyy Finereader* (many versions, also server-based)
  - Omnireader
  - Others..

# 5. Seeing is believing

- What is the **real problem** here?

5 + LPS 7.258

Ba 6.3
Bd 5.3
Hb 1.3
Hb 1.32
Hb 1.35

B R I T T E N , Benjamin

1) 2'45 1-4. Folksong arrangements, vol.I: British Isles (Kansanlaulu-
2) 2'15 sovituksia Brittien saarilta)
3) 2'17 1. The Sally Gardens (n:o 1) (Rantapuistossa) .-
4) 0'37 2. The bonny earl o'Moray(n:o 3)(Morayn jaarli) .-
3. The ash grove (n:o 6) (Saarnilehto) .-
4. Oliver Cromwell (n:o 7) .-
5-12) 5-12. Winter words, op. 53, laulusarja Thomas Hardyn runoihin ja
balladeihin (Talvisia säkeitä) .- //
13-21 13-21. The Holy Sonnets of John Donne,op. 35, laulusarja.
24'37

G i l v a n , Raimund (ten); Capon,Frederic (pf)

1,2) DA CAMERA SM 90002 ; 1966; esleht. ra + tks eng ; 12 + 9 r
3,4) MUSICA RARA MUS 25 ;

RK/ aa kpl 1,2,3,4 OK

6 pk + 6 ek + 16 syl = 35 pun,
10 val

# 5. Seeing is believing

KAR

- What is the **real problem** here?

+LPS 7.262
(levypuoli 1)

O O P P E R A-A A R I O I T A  S A K S A N K I E L E L L Ä

1) 1'22
2) 1'45
3) 2'55
4) 4'34
5) 2'52
6) 2'35
7) 6'02

1. Mozart: Don Giovanni: I: Finch' han dal vino (Don Giovannin sampanja-aaria). - 2. Mozart: Don Giovanni: II: Deh vieni alla finestra (Don Giovannin serenadi). - 3. Mozart: Don Giovanni: II: Metà di voi qua vadano (Don Giovannin aaria). - 4. Mozart: Le Nozze di Figaro: III: Hai gia vinto la causa...Vedro, mentr'io sospiro (Kreivi Almavivan aaria). - 5. Beethoven: Fidelio: I: Ha! welch ein Augenblick! (Pizarron aaria). - 6. Verdi: Un ballo in maschera: I: Alla vita che t'arride (Renaton aaria). - 7. Verdi: Un ballo in maschera: III: Alzati!...Eri tu (Renaton resit: & aaria). - //

F i s c h e r-D i e s k a u, Dietrich (bar); RSO Berlin (1-3); Berliner Philharmoniker (4); Bayerischea Staatsorchester (5); Orch. der Deutschen Oper Berlin (6); L ö w l e i n, Hans (joht-1-3); L e i t n e r, Ferdinand (4); F r i c s a y, Ferenc (joht-5); P a t a n e, Giuseppe (joht-6,7); Bayerischer Staats-opernchor (5).

RK/tp

DGG SLPEM 136460;1964; sel. sa takak.; 7 + (4)r
kpl 1 7pk+1ek - 15 pun 4 val

# 5. Seeing is believing

- What is the **real problem** here?

5 + LPS 7.258

Ba 6.3
Bd 5.3
Hb 1.3 B R I T T E N , Benjamin
Hb 1.32
Hb 1.35 1-4. Folksong arrangements, vol.I: British Isles (Kansanlaulu-
sovituksia Brittien saarilta)
1) 2'45 1. The Sally Gardens (n:o 1) (Rantapuistossa) .-
2) 2'15 2. The bonny earl o'Moray(n:o 3)(Morayn jaarli) .-
3) 2'17 3. The ash grove (n:o 6) (Saarnilehto) .-
4) 0'37 4. Oliver Cromwell (n:o 7) .-
5-12) 5-12. Winter words, op. 53, laulusarja Thomas Hardyn runoihin ja
balladeihin (Talvisia säkeitä) .- //
13-21. The Holy Sonnets of John Donne,op. 35, laulusarja.

G i l v a n , Raimund (ten); Capon,Frederic (pf)

1,2) DA CAMERA SM 90002 ; 1966; esleht. ra + tks eng ; 12 + 9 r
3,4) MUSICA RARA MUS 25 ;

RK/ aa kpl 1,2,3,4 OK 6 pk + 6 ek + 16 syl = 35 pun,
10 val

+LPS 7.262
(levypuoli 1)

O O P P E R A-A A R I O I T A S A K S A N K I E L E L L Ä

1) 1'22 1. Mozart: Don Giovanni: I: Finch' han dal vino (Don Giovannin
samppanja-aaria). - 2. Mozart: Don Giovanni: II: Deh vieni alla
2) 1'45 finestra (Don Giovannin serenadi). - 3. Mozart: Don Giovanni:
II: Metà di voi qua vadano (Don Giovannin aaria). - 4. Mozart:
3) 2'55 Le Nozze di Figaro: III: Hai gia vinto la causa...Vedro, mentr'io
sospiro (Kreivi Almavivan aaria). - 5. Beethoven: Fidelio: I:
4) 4'34 Ha! welch ein Augenblick! (Pizarron aaria). - 6. Verdi: Un ballo
in maschera: I: Alla vita che t'arride (Renaton aaria). -
5) 2'52 7. Verdi: Un ballo in maschera: III: Alzati!...Eri tu (Renaton
resit: & aaria). - //
6) 2'35
F i s c h e r-D i e s k a u, Dietrich (bar); RSO Berlin (1-3);
7) 6'02 Berliner Philharmoniker (4); Bayerischea Staatsorchester (5);
Orch. der Deutschen Oper Berlin (6); L ö w l e i n, Hans (joht-
1-3); L e i t n e r, Ferdinand (4); F r i c s a y, Ferenc (joht-5);
P a t a n e, Giuseppe (joht-6,7); Bayerischer Staats-
opernchor (5).

RK/tp DGG SLPEM 136460;1964; sel. sa takak.; 7 + (4)r
kpl 1 7pk+1ek = 15 pun 4 val

# 0+1. I said: take the FREE LUNCH!

KAR

- If somebody wants to offer you a free lunch, take it!
  - We have 37 000 cardboards from our vinyl archives, ca. 1966 – 1973. Care to scan & **OCR** them, as nobody else cannot do it?

They are, of course, **all different** (as such a computer would regard them)!

# 5. Seeing is believing

KAR

- Even without extensive knowledge about OCR, one can understand that

  **OCR itself should not be the biggest problem**

  - *Machine written text* (special typewriter types of machines; 2 or 3 machines making slightly different typeface)
    - the OCR should not be the major bottleneck, at least when using **well-trained** (to the actual cards) & **able OCR software**

# 5. Seeing is believing

KAR

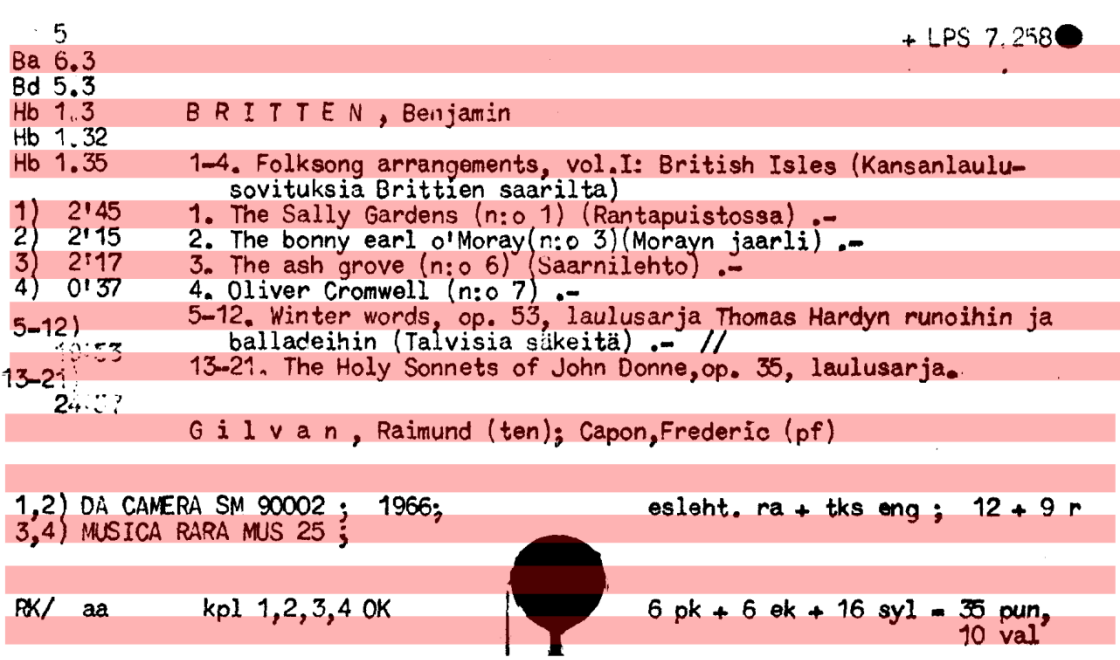
- OCR will **always** have problems that are inherent to the originals
  - Some of the cards are pressed each other and/or otherwise worn (*bad condition originals)*
    - the OCR will make random mistakes there, no doubt about it (but their total amount is relatively small)
  - Some of the cards include *machine and/or hand-made corrections* (e.g., double strikes; *hand-written originals*)
    - the OCR will make mistakes there, no doubt about it (but their total amount is small)

# 5. Seeing is believing

KAR

- But the layout in the cards is the PROBLEM
  - The information in the cards overlaps in visible rows , which is
    **A REAL PROBLEM**
  - AND the *variance in the actual layout between the cards* is
    **EVEN MORE SERIOUS DILEMMA**
- *Technical vs. practical problems*

# 5. Seeing is believing



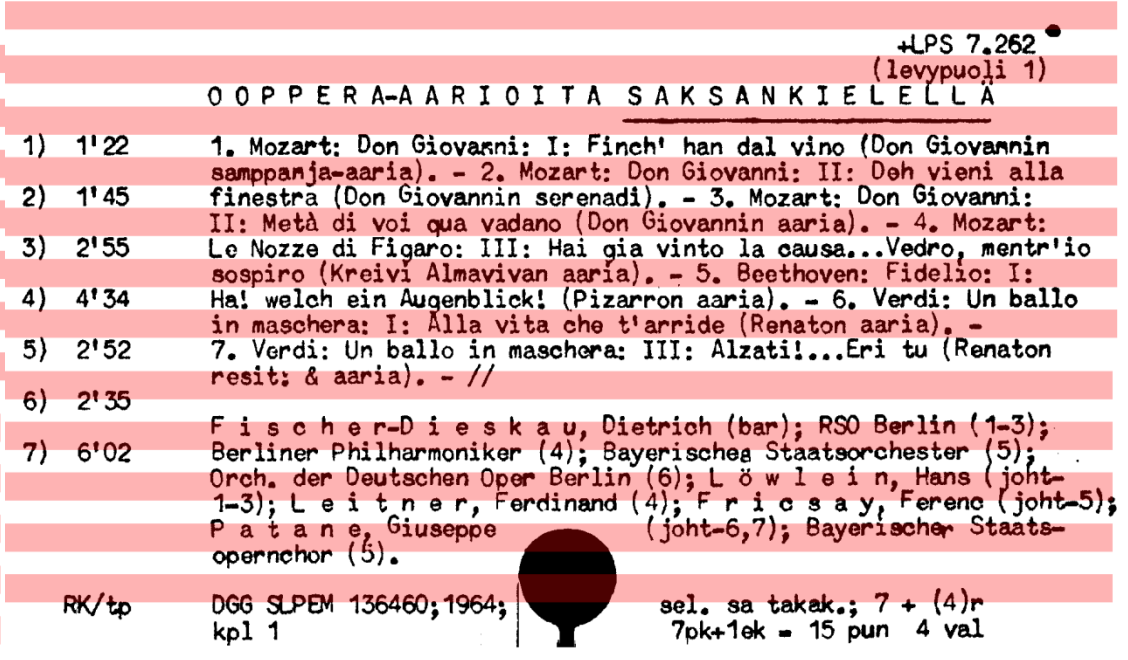
5 + LPS 7.258

Ba 6.3
Bd 5.3
Hb 1.3 B R I T T E N , Benjamin
Hb 1.32
Hb 1.35 1-4. Folksong arrangements, vol.I: British Isles (Kansanlaulu-
sovituksia Brittien saarilta)
1) 2'45 1. The Sally Gardens (n:o 1) (Rantapuistossa) .-
2) 2'15 2. The bonny earl o'Moray(n:o 3)(Morayn jaarli) .-
3) 2'17 3. The ash grove (n:o 6) (Saarnilehto) .-
4) 0'37 4. Oliver Cromwell (n:o 7) .-
5-12) 5-12. Winter words, op. 53, laulusarja Thomas Hardyn runoihin ja
balladeihin (Talvisia säkeitä) .- //
13-21) 13-21. The Holy Sonnets of John Donne,op. 35, laulusarja.

G i l v a n , Raimund (ten); Capon,Frederic (pf)

1,2) DA CAMERA SM 90002 ; 1966; esleht. ra + tks eng ; 12 + 9 r
3,4) MUSICA RARA MUS 25 ;

RK/ aa kpl 1,2,3,4 OK 6 pk + 6 ek + 16 syl = 35 pun,
10 val



+LPS 7.262
(levypuoli 1)

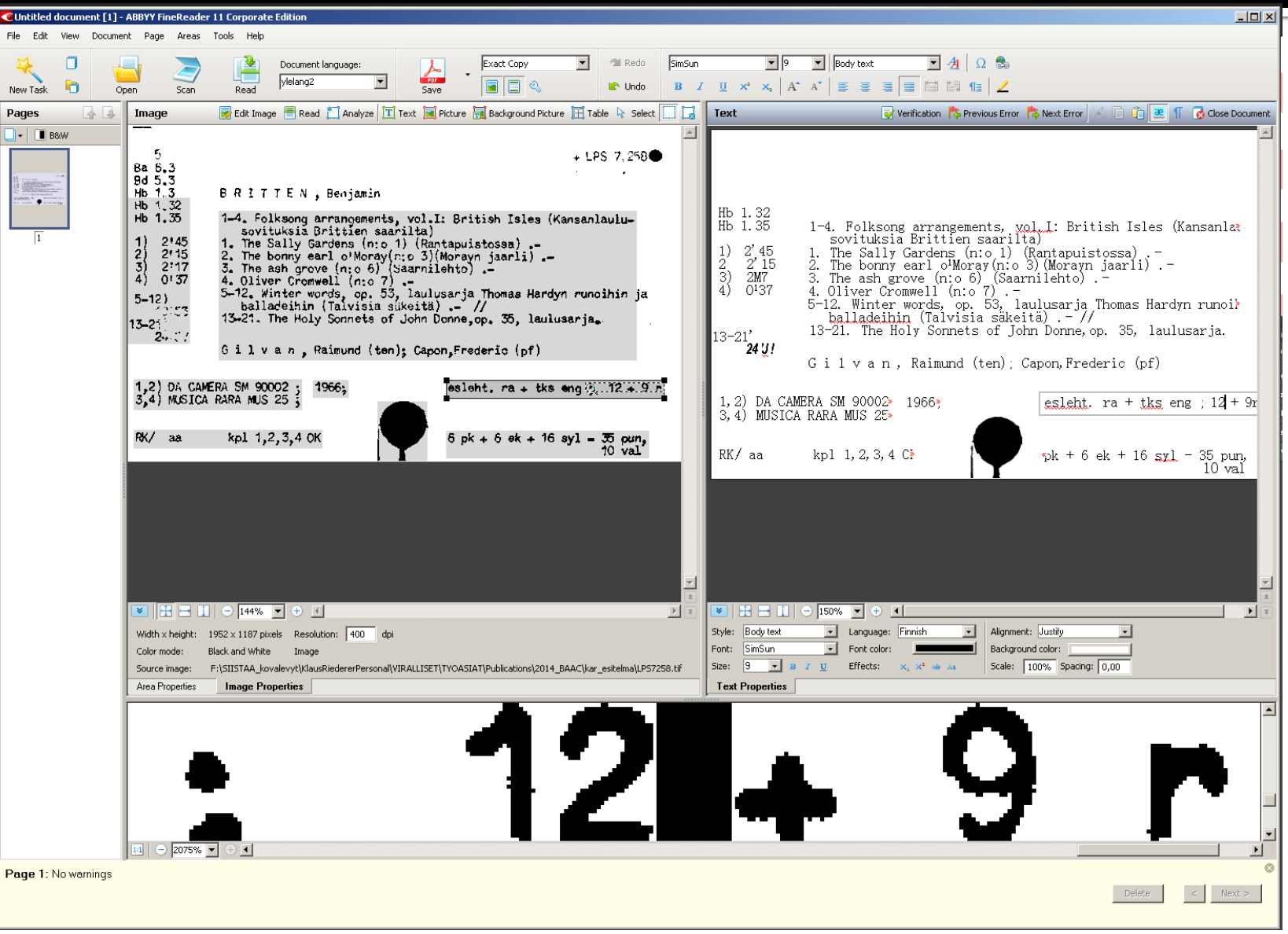
O O P P E R A-A A R I O I T A S A K S A N K I E L E L L Ä

1) 1'22 1. Mozart: Don Giovanni: I: Finch' han dal vino (Don Giovannin samppanja-aaria). - 2. Mozart: Don Giovanni: II: Deh vieni alla
2) 1'45 finestra (Don Giovannin serenadi). - 3. Mozart: Don Giovanni:
II: Metà di voi qua vadano (Don Giovannin aaria). - 4. Mozart:
3) 2'55 Le Nozze di Figaro: III: Hai gia vinto la causa...Vedro, mentr'io
sospiro (Kreivi Almavivan aaria). - 5. Beethoven: Fidelio: I:
4) 4'34 Ha! welch ein Augenblick! (Pizarron aaria). - 6. Verdi: Un ballo
in maschera: I: Alla vita che t'arride (Renaton aaria). -
5) 2'52 7. Verdi: Un ballo in maschera: III: Alzati!...Eri tu (Renaton
resit: & aaria). - //
6) 2'35
F i s c h e r-D i e s k a u, Dietrich (bar); RSO Berlin (1-3);
7) 6'02 Berliner Philharmoniker (4); Bayerischea Staatsorchester (5);
Orch. der Deutschen Oper Berlin (6); L ö w l e i n, Hans (joht-
1-3); L e i t n e r, Ferdinand (4); F r i c s a y, Ferenc (joht-5);
P a t a n e, Giuseppe (joht-6,7); Bayerischer Staats-
opernchor (5).

RK/tp DGG SLPEM 136460;1964; sel. sa takak.; 7 + (4)r
kpl 1 7pk+1ek = 15 pun 4 val

- The information in the cards overlaps in visible rows , which is **A REAL PROBLEM**
- AND the *variance in the actual layout between the cards* is **EVEN MORE SERIOUS DILEMMA**

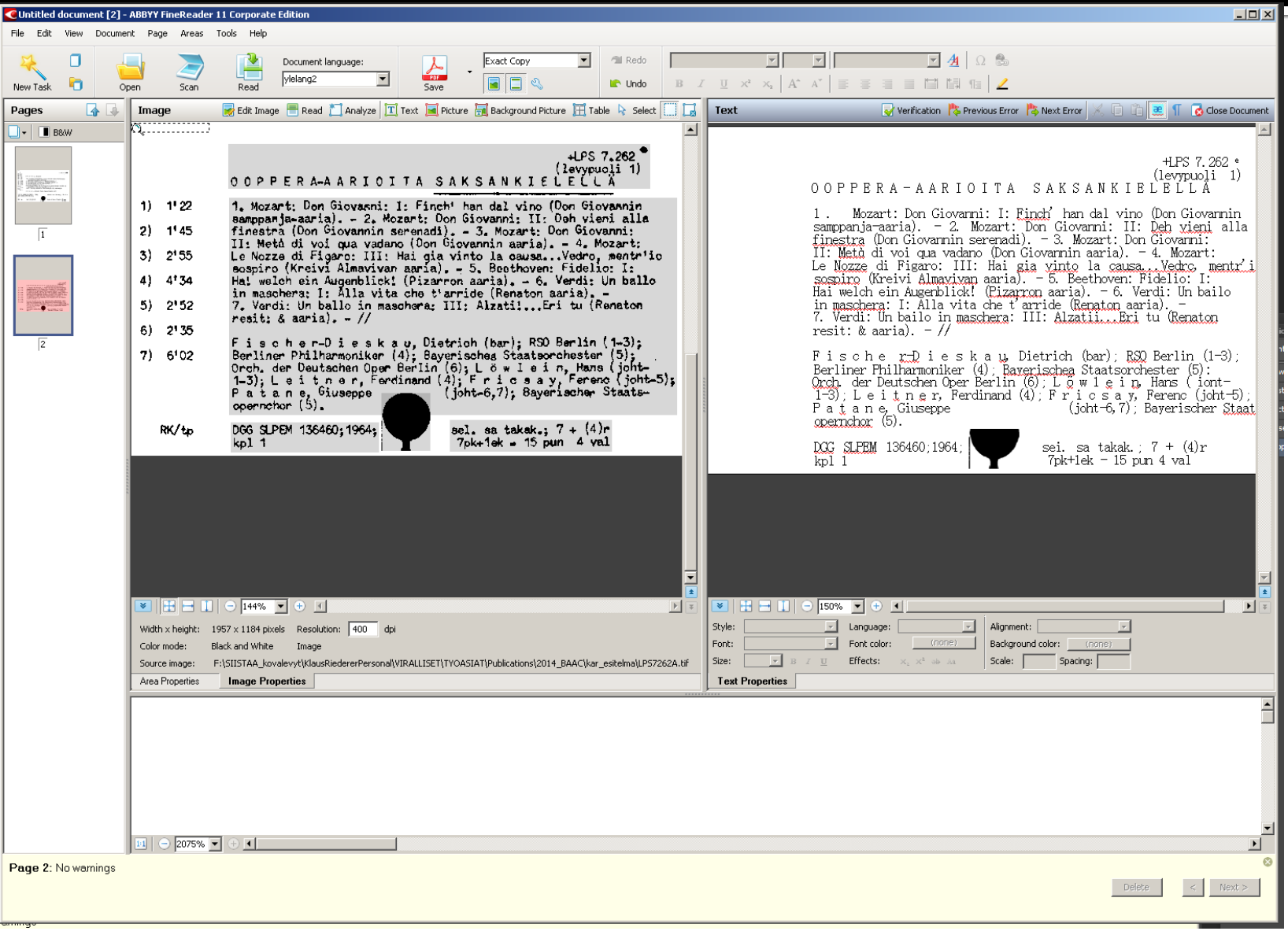
# 6. Standard OCR




- Yes, OCR software can make OCR read areas (regions) of the scanned image
  - but how well does it work?

KAR

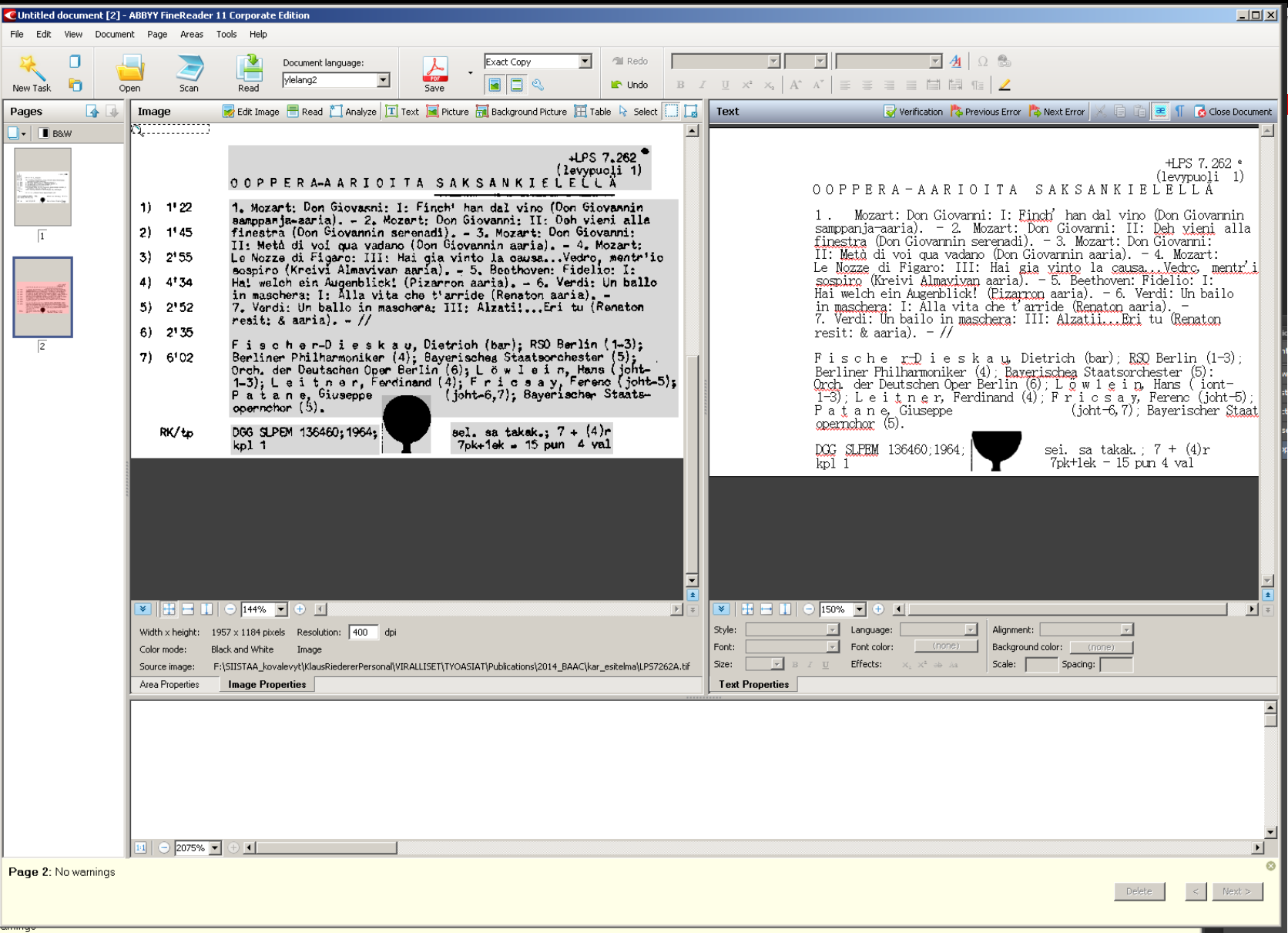
# 6. Standard OCR does




- Yes, OCR software can make OCR read areas (regions) of the scanned image
  - but how well does it work?

KAR

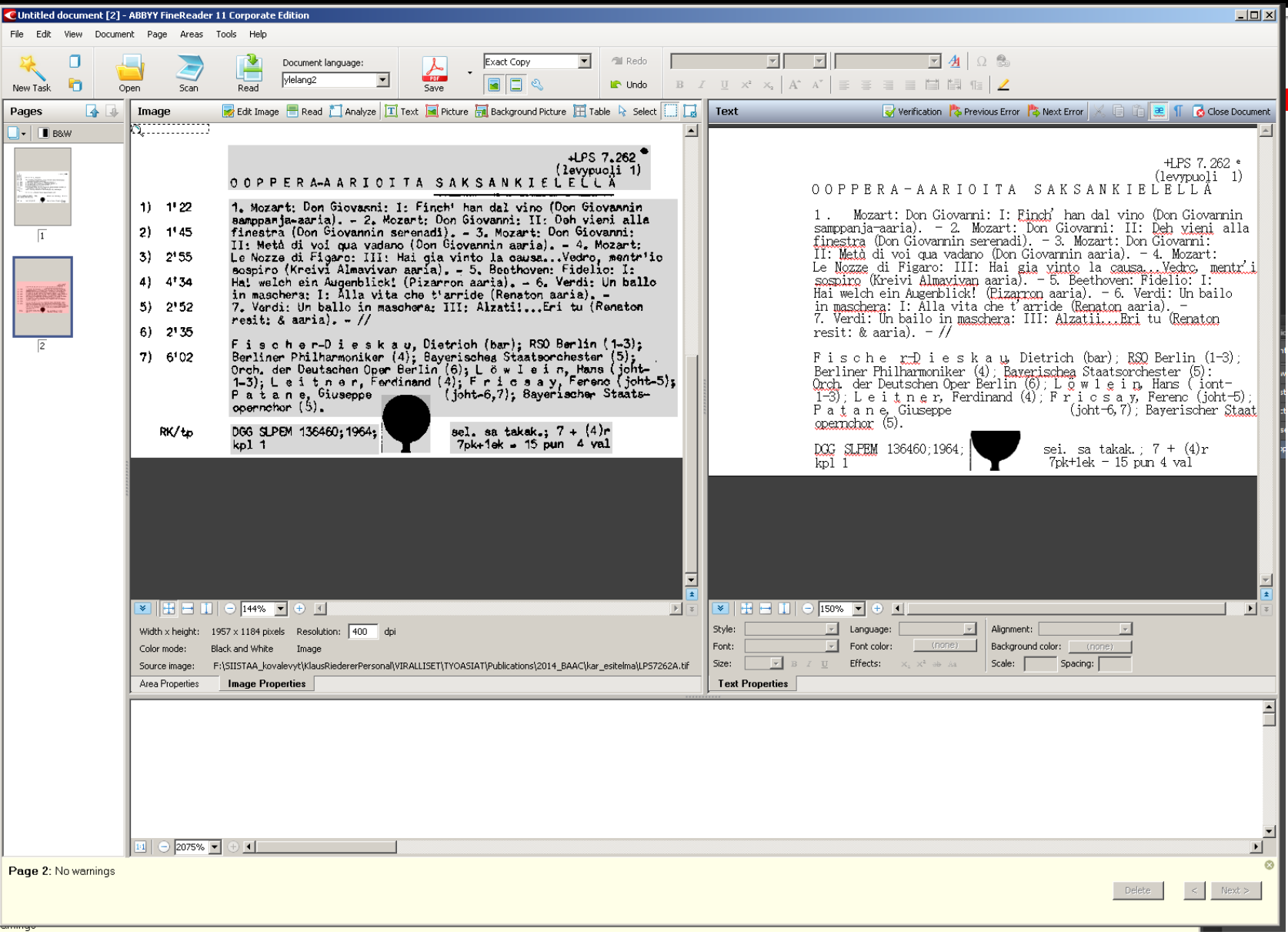
# 6. Standard OCR does not




Yes, OCR software can make OCR read areas (regions) of the scanned image

- but how well does it work?

# 6. Standard OCR does not work




- Yes, you can use in OCR software preset OCR read areas (regions)
  - but what if all cards are different?

# 6. Standard OCR does not work!

+LPS 7.262
(levypuoli 1)

OOPPERA-AARIOITA SAKSANKIELELLÄ

1) 1'22
2) 1'45
3) 2'55
4) 4'34
5) 2'52
6) 2'35
7) 6'02

1. Mozart: Don Giovanni: I: Finch' han dal vino (Don Giovannin sampanja-aaria). - 2. Mozart: Don Giovanni: II: Deh vieni alla finestra (Don Giovannin serenadi). - 3. Mozart: Don Giovanni: II: Metä di voi qua vadano (Don Giovannin aaria). - 4. Mozart: Le Nozze di Figaro: III: Hai gia vinto la causa...Vedro, mentr'io sospiro (Kreivi Almavivan aaria). - 5. Beethoven: Fidelio: I: Ha! welch ein Augenblick! (Pizarron aaria). - 6. Verdi: Un ballo in maschera: I: Alla vita che t'arride (Renaton aaria). - 7. Verdi: Un ballo in maschera: III: Alzati!...Eri tu (Renaton resit: & aaria). - //

F i s c h e r-D i e s k a u, Dietrich (bar); RSO Berlin (1-3); Berliner Philharmoniker (4); Bayerisches Staatsorchester (5); Orch. der Deutschen Oper Berlin (6); L ö w l e i n, Hans (joht-1-3); L e i t n e r, Ferdinand (4); F r i c s a y, Ferenc (joht-5); P a t a n e, Giuseppe (joht-6,7); Bayerischer Staats-opernchor (5).

RK/tp

DGG SLPEM 136460;1964; kpl 1

sel. sa takak.; 7 + (4)r 7pk+1ek = 15 pun 4 val

- Do the region splitting manually per card?
  - And split your brains while doing it (37000*8 = 296 000 regions)?!

# 7. PROBLEM – ~~what~~ how to do?

KAR

- Only one (obvious) solution here:
  - split the scanned image into regions and OCR those separately
  - **HOW TO MAKE THIS EFFECTIVELY?**

# 7. KAR Ltd. = problem solver

KAR

- KAR Ltd / Klaus A J Riederer, PhD:
  +20 years of coding – now applying also *machine vision* methods
  - Divide the scanned card image into 8 regions, applying deep intelligence *mimicking human cognition*
  - Each card is *processed twice*, both processes use *multiple inter-dependent functioning levels*

# 7. KAR Ltd.'s problem solving

KAR

- *The first run applies preset region positions (i.e., **region of interest =ROI**) for machine vision methods*
  - *Computer suggests the most probable region area for the user, based on numerous boundary conditions and preset rules*
  - *The machine vision methods and its conditions and rules are fine-tuned very carefully, separately for each region in question*

# 7. KAR Ltd.'s problem solving

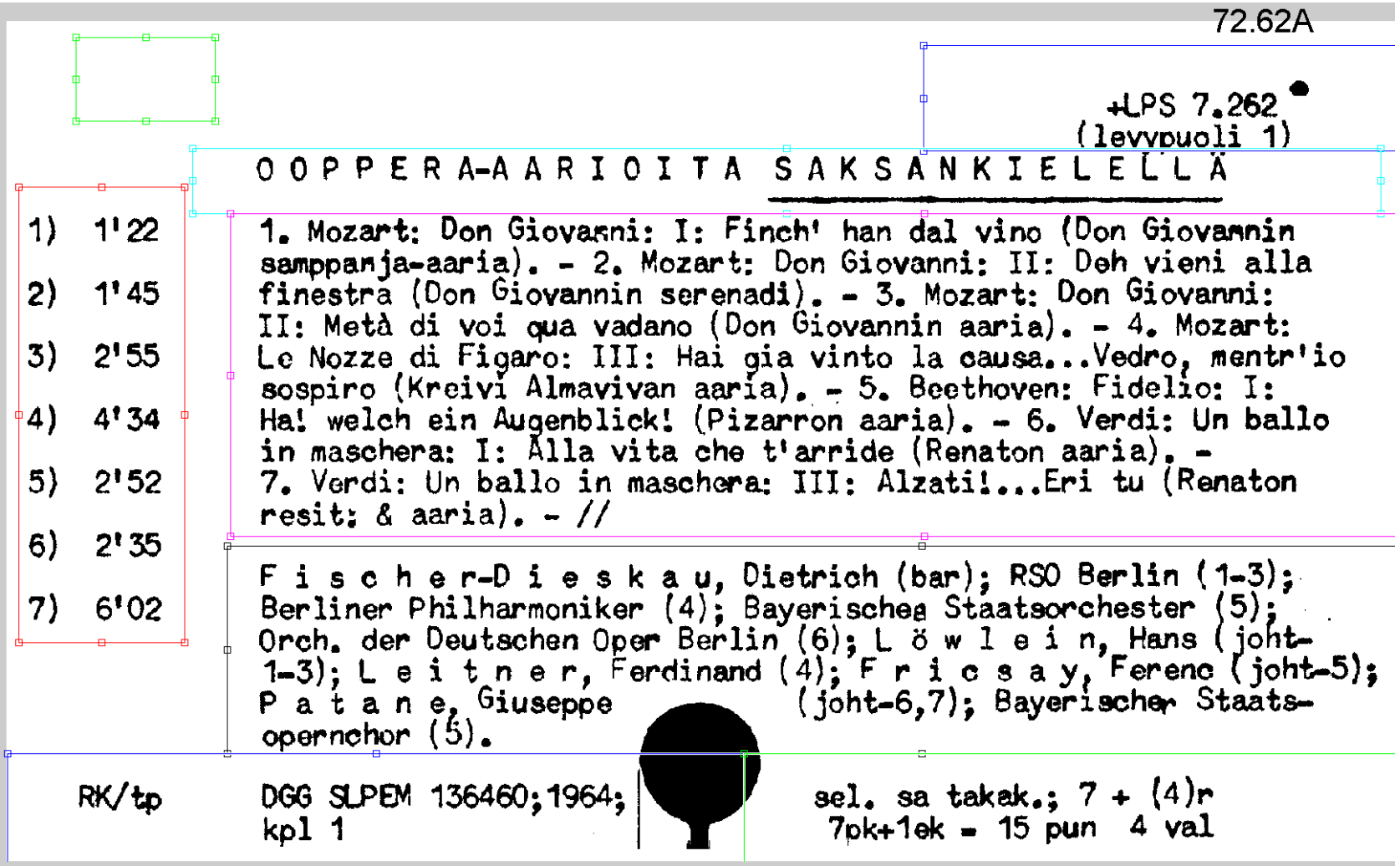
KAR

- *In the second run the user verifies and accepts or changes the region positions from the first run, because:*
  - *The computer "learns" from the user's response so that following regions' positions depend on the previous regions' positions (<= regions may not overlap!)*
    - *This utilization of history is made only per current card, as the cards are so varying between each other*
    - *A global learning process is inherent in the coding process: the code has between iteratively improved (until ca. 15000 cards where run)*

# 7. KAR Ltd.'s problem solving

KAR

- *Divide the image into 8 position-varying regions*

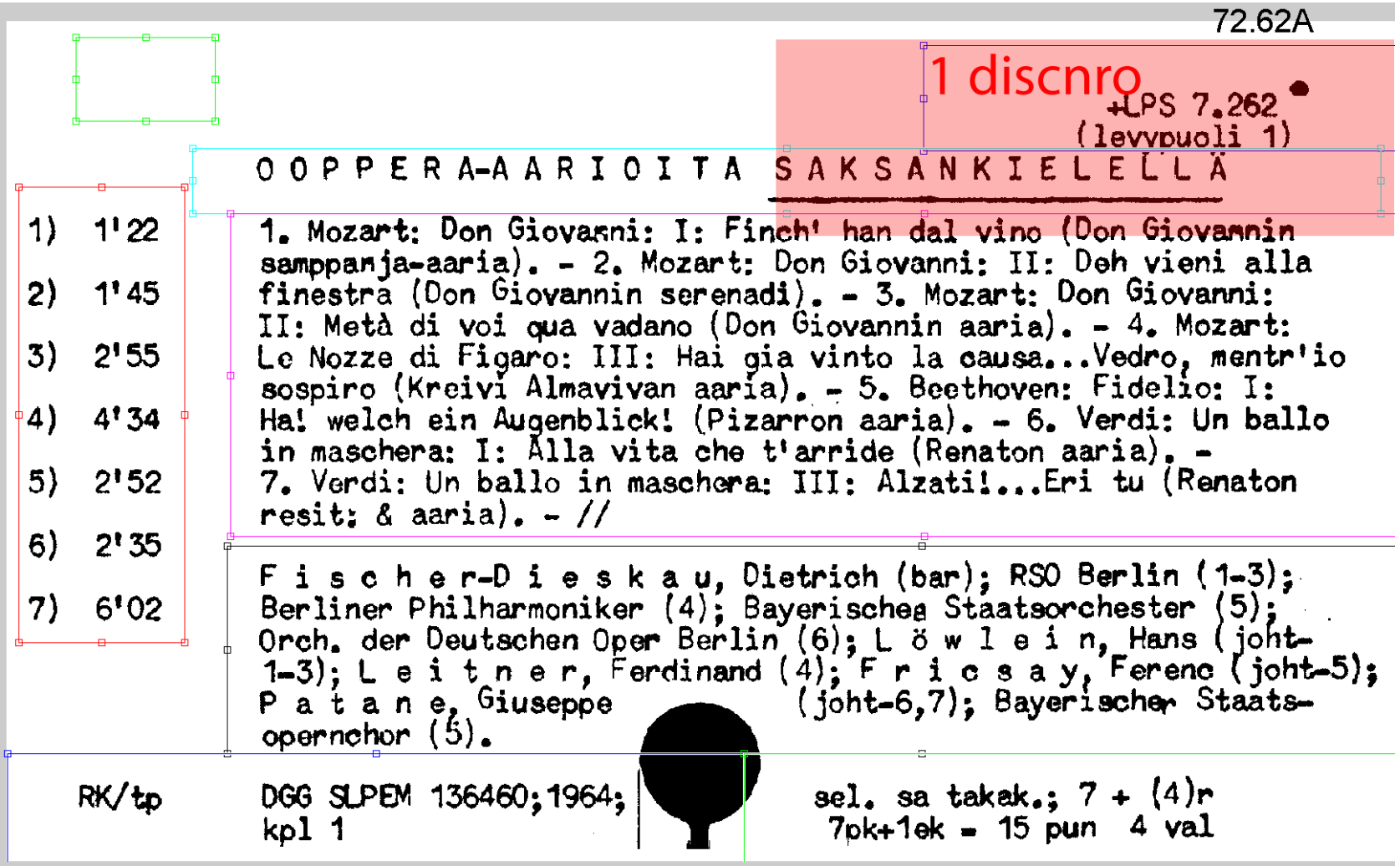



1 discnro ("disc number")
2 class ("classification")
3 songlen ("song length")
4 title ("disc title")
5 songname ("name per song")
6 artist ("artist per disc")
7 publisher ("disc publisher")
8 addinfo ("additional information")

# 7. KAR Ltd. problem solving

KAR

- *Divide the image into 8 position-varying regions*

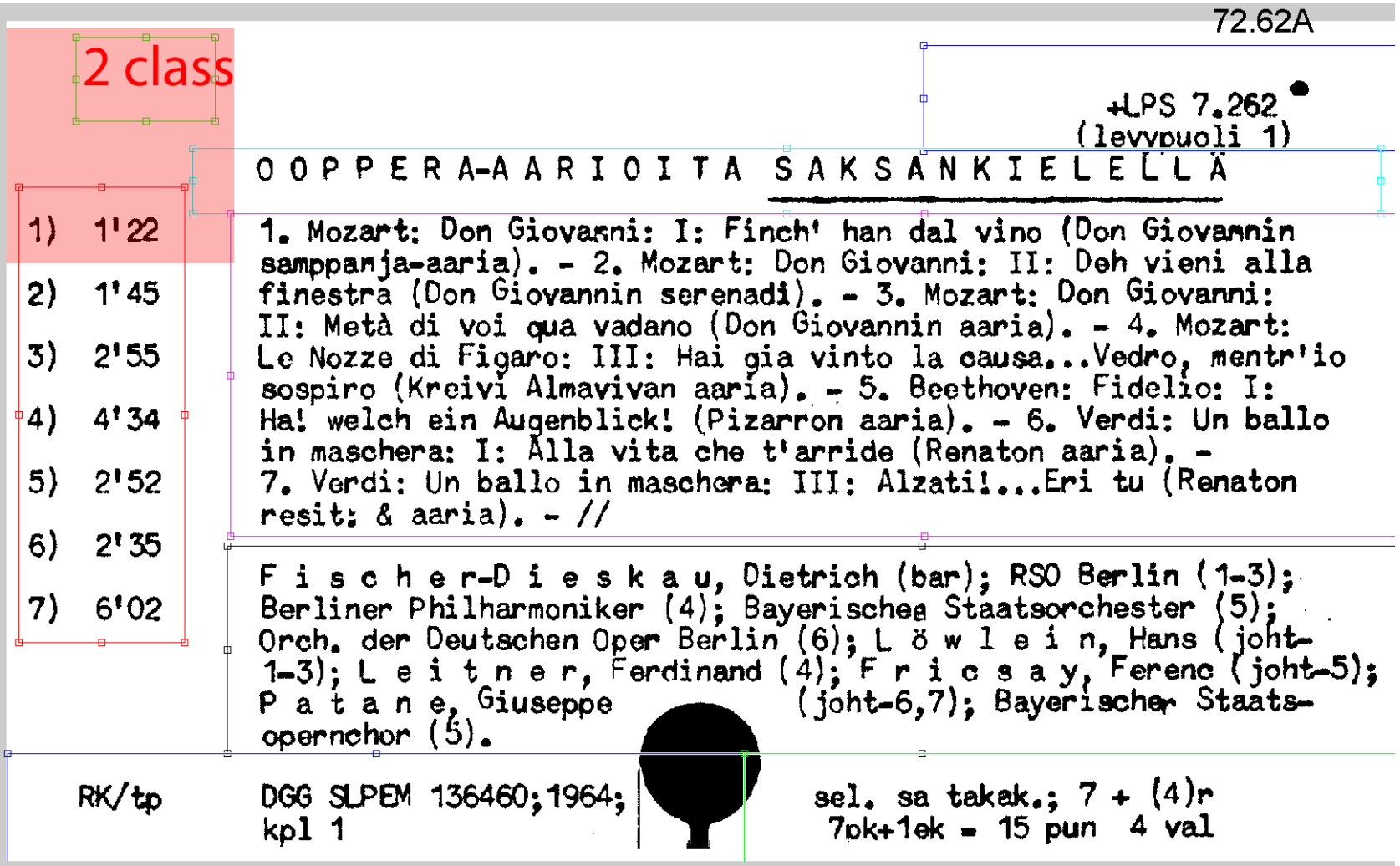



1 discnro ("disc number")
2 class ("classification")
3 songlen ("song length")
4 title ("disc title")
5 songname ("name per song")
6 artist ("artist per disc")
7 publisher ("disc publisher")
8 addinfo ("additional information")

# 7. KAR Ltd.'s problem solving

KAR

- *Divide the image into 8 position-varying regions*

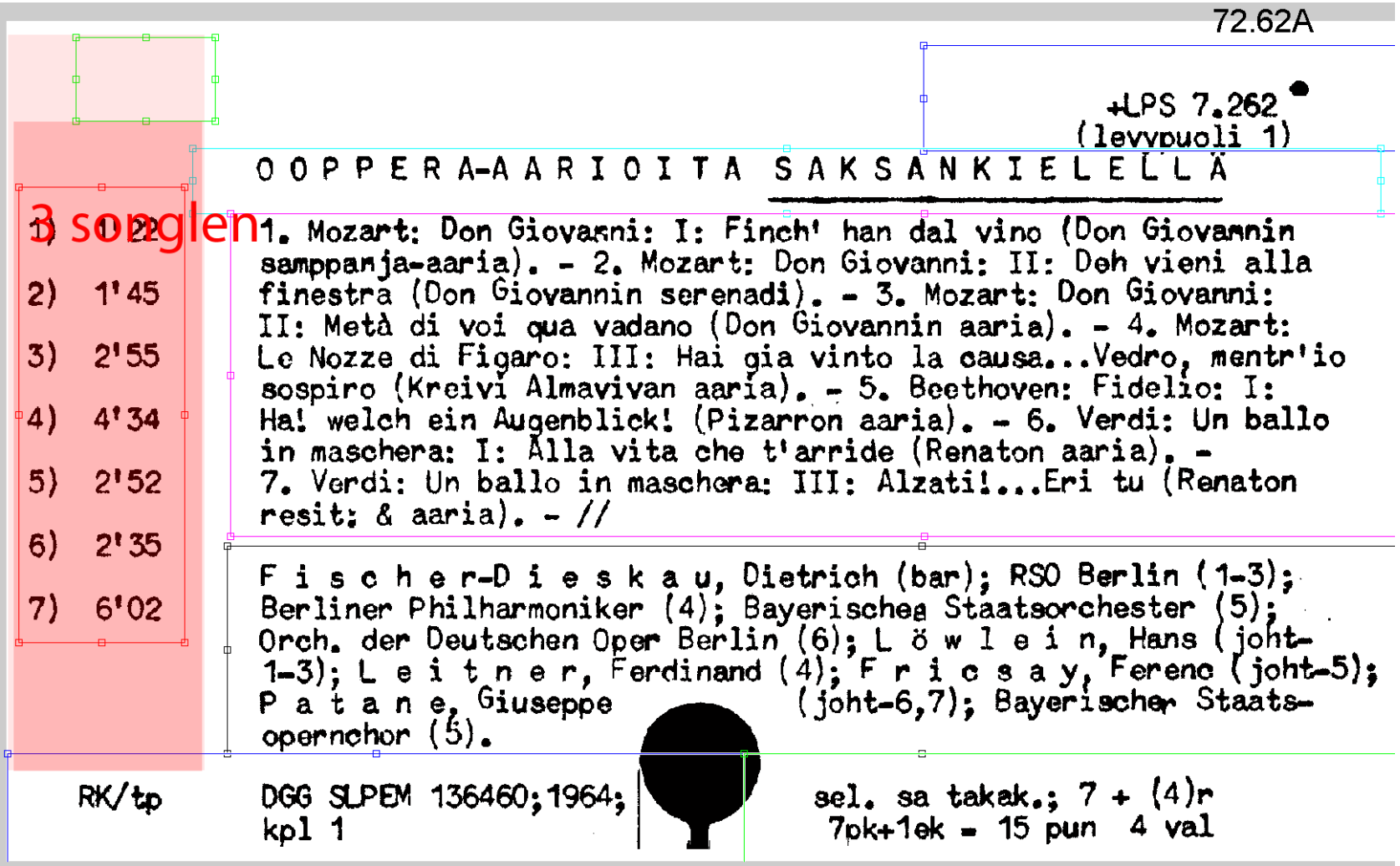



1 discnro ("disc number")
2 class ("classification")
3 songlen ("song length")
4 title ("disc title")
5 songname ("name per song")
6 artist ("artist per disc")
7 publisher ("disc publisher")
8 addinfo ("additional information")

# 7. KAR Ltd.'s problem solving

KAR

- *Divide the image into 8 position-varying regions*

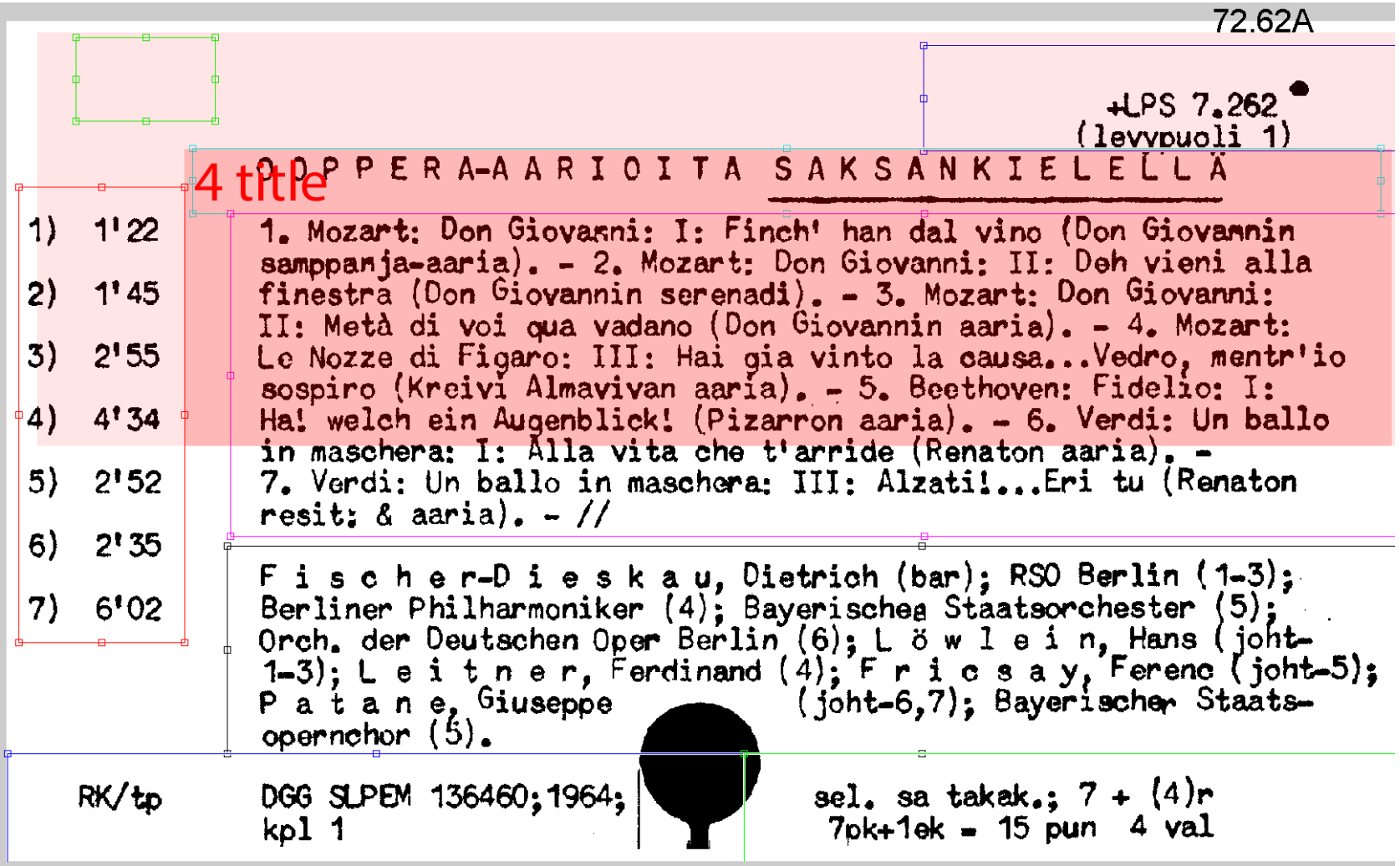



1 discnro ("disc number")
2 class ("classification")
3 songlen ("song length")
4 title ("disc title")
5 songname ("name per song")
6 artist ("artist per disc")
7 publisher ("disc publisher")
8 addinfo ("additional information")

# 7. KAR Ltd. 's problem solving

KAR

- *Divide the image into 8 position-varying regions*

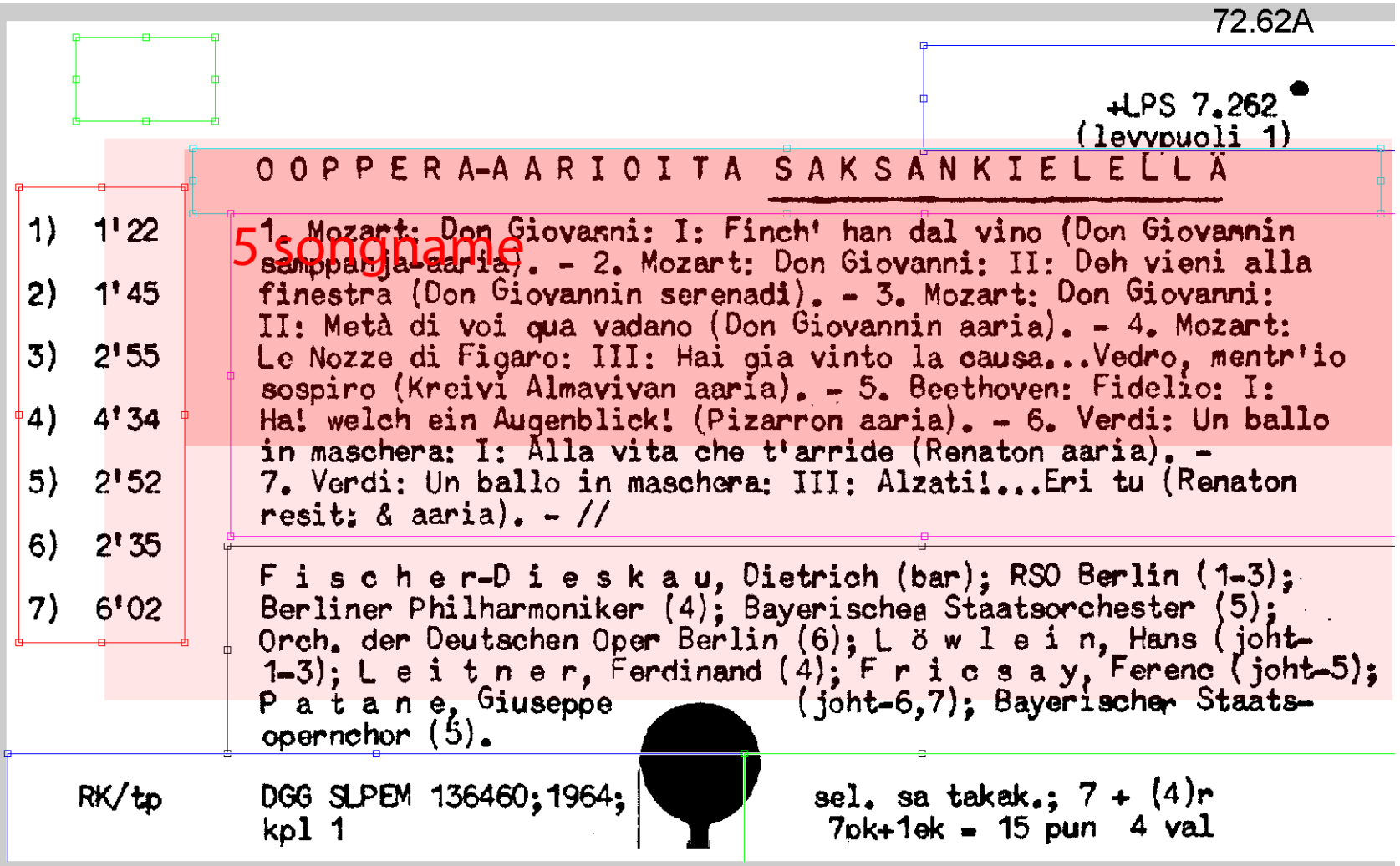



1 discnro ("disc number")
2 class ("classification")
3 songlen ("song length")
4 title ("disc title")
5 songname ("name per song")
6 artist ("artist per disc")
7 publisher ("disc publisher")
8 addinfo ("additional information")

# 7. KAR Ltd.'s problem solving

KAR

- *Divide the image into 8 position-varying regions*

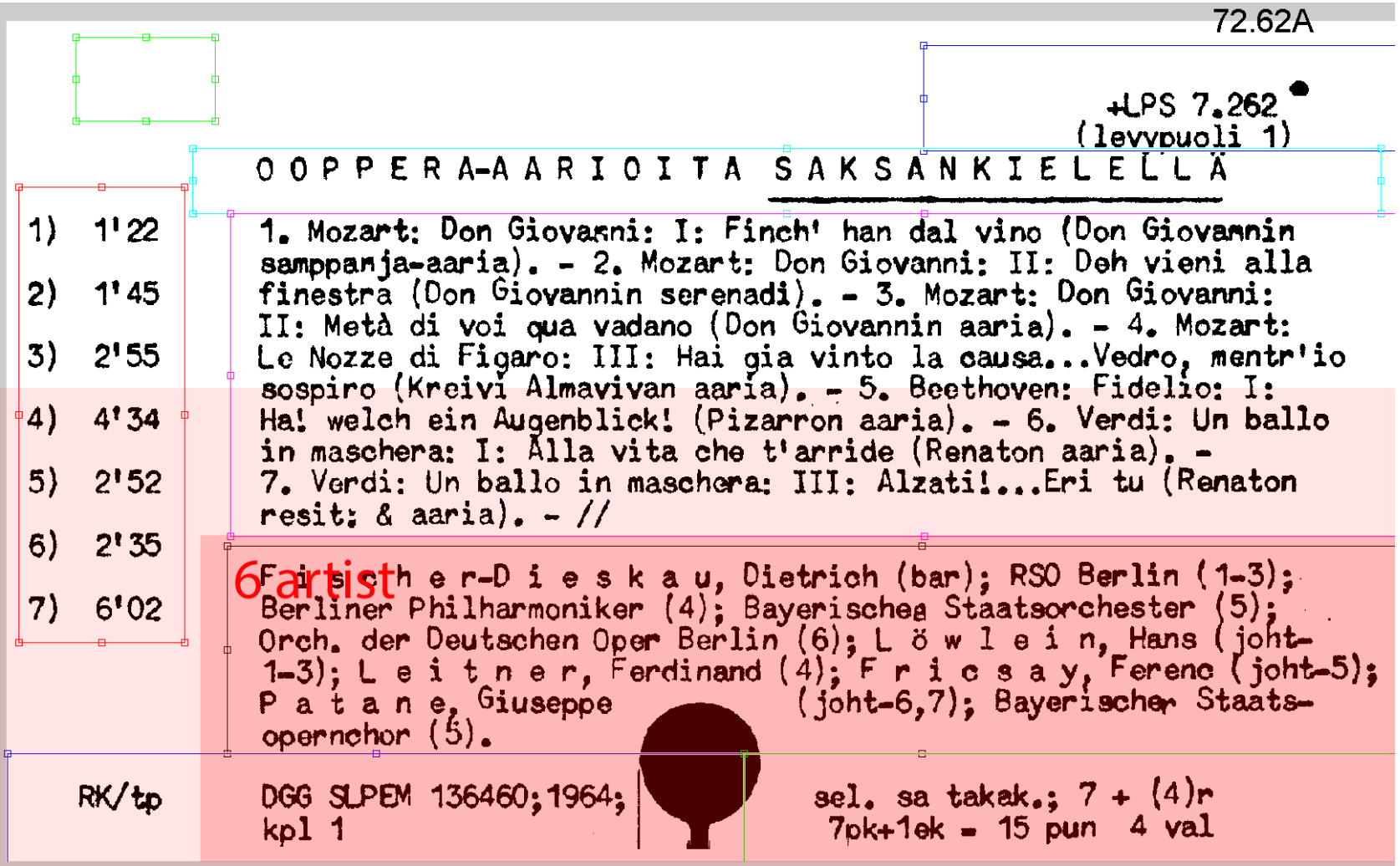



1 discnro ("disc number")
2 class ("classification")
3 songlen ("song length")
4 title ("disc title")
5 songname ("name per song")
6 artist ("artist per disc")
7 publisher ("disc publisher")
8 addinfo ("additional information")

# 7. KAR Ltd.'s problem solving

KAR

- *Divide the image into 8 position-varying regions*

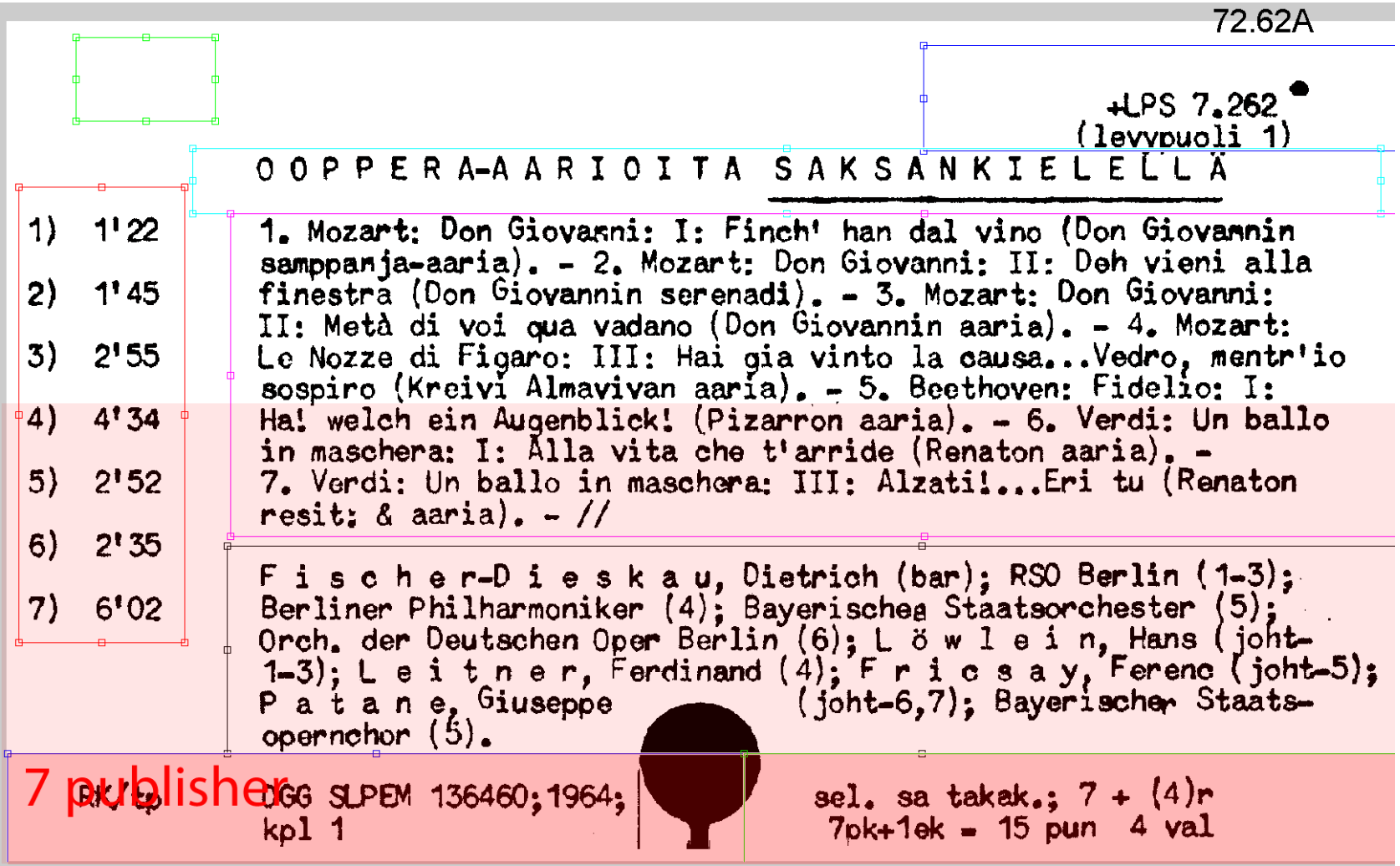



1 discnro ("disc number")
2 class ("classification")
3 songlen ("song length")
4 title ("disc title")
5 songname ("name per song")
6 artist ("artist per disc")
7 publisher ("disc publisher")
8 addinfo ("additional information")

# 7. KAR Ltd.'s problem solving

KAR

- *Divide the image into 8 position-varying regions*

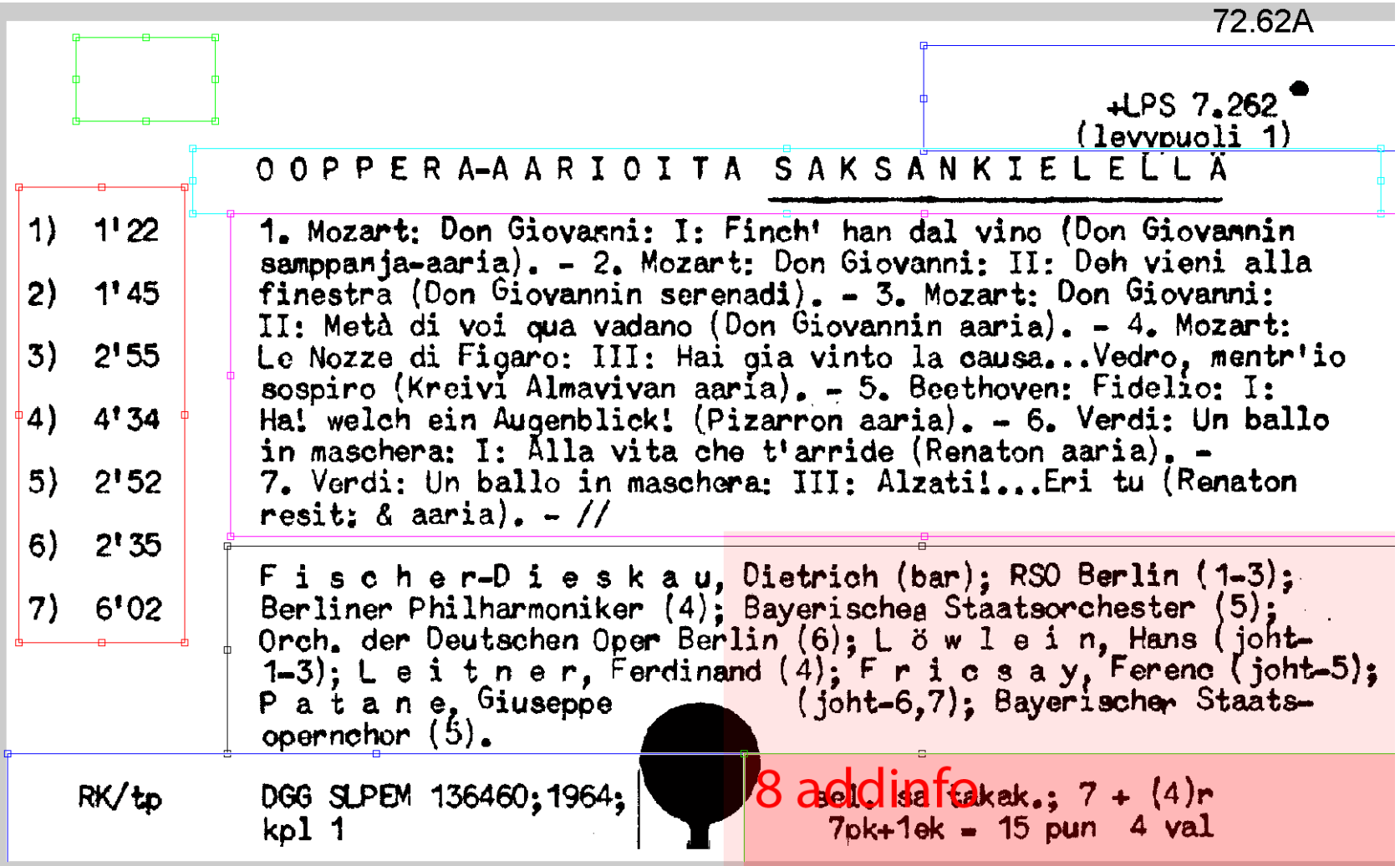



1 discnro ("disc number")
2 class ("classification")
3 songlen ("song length")
4 title ("disc title")
5 songname ("name per song")
6 artist ("artist per disc")
7 publisher ("disc publisher")
8 addinfo ("additional information")

# 7. KAR Ltd.'s problem solving

KAR

- *Divide the image into 8 position-varying regions*




1 discnro ("disc number")
2 class ("classification")
3 songlen ("song length")
4 title ("disc title")
5 songname ("name per song")
6 artist ("artist per disc")
7 publisher ("disc publisher")
8 addinfo ("additional information")

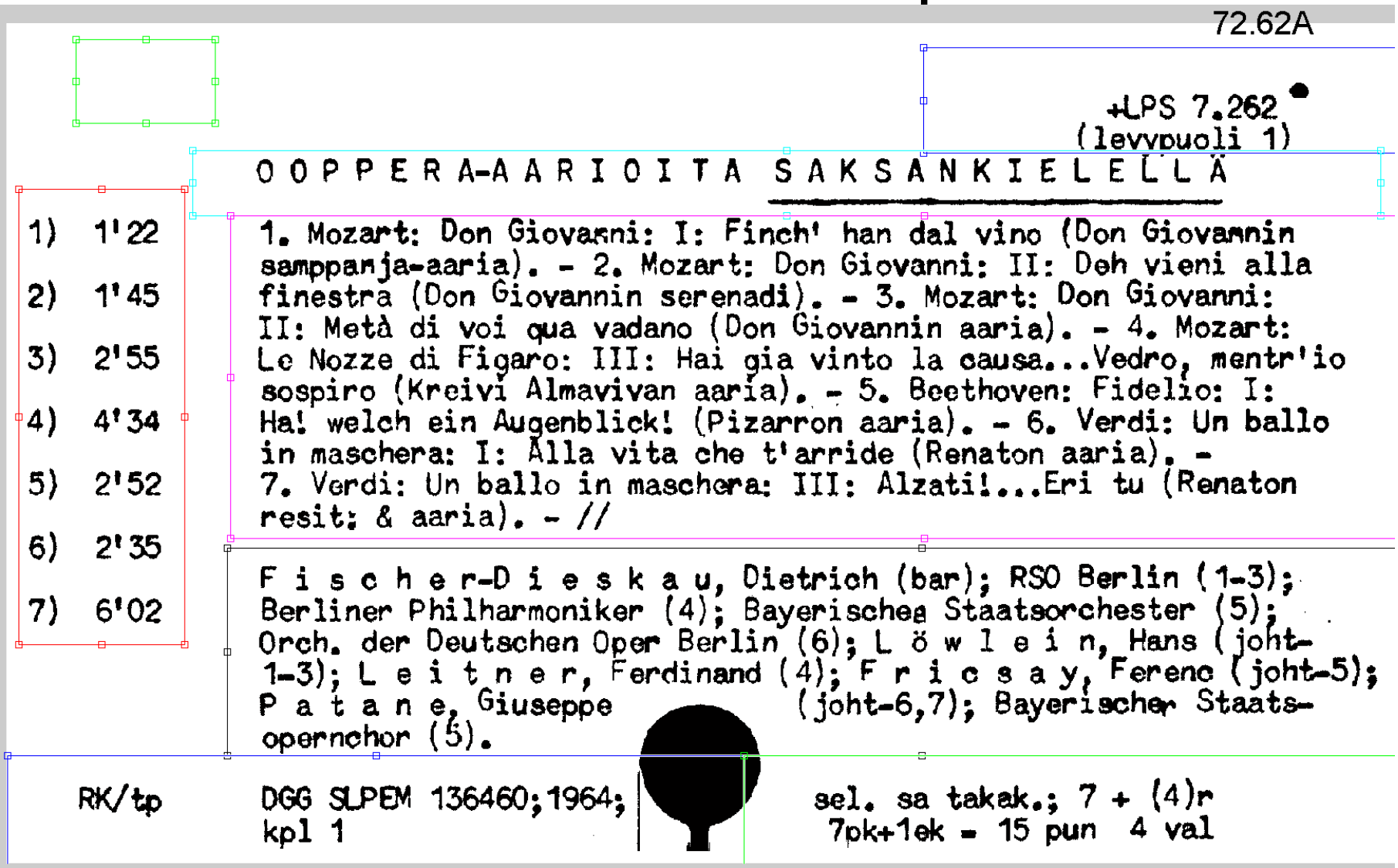
# 7. KAR Ltd.'s problem solving

- Most basic rules for the regions:
  - All the regions are rectangle –shaped
  - The card must have ONLY information from the region #1 (discnro), BUT other region information may be missing (= empty)
  - Some of the region positions do not overlap; but
  - The final region areas may overlap (so that all information is gathered from the cards, in cost of redundancy)

# 7. KAR Ltd.'s problem solving

KAR

- How does this work in practice?



72.62A

+LPS 7.262
(levypuoli 1)

O O P P E R A-A A R I O I T A  S A K S A N K I E L E L L Ä

1) 1'22
2) 1'45
3) 2'55
4) 4'34
5) 2'52
6) 2'35
7) 6'02

1. Mozart: Don Giovanni: I: Finch' han dal vino (Don Giovannin sampanja-aaria). - 2. Mozart: Don Giovanni: II: Deh vieni alla finestra (Don Giovannin serenadi). - 3. Mozart: Don Giovanni: II: Metà di voi qua vadano (Don Giovannin aaria). - 4. Mozart: Le Nozze di Figaro: III: Hai gia vinto la causa...Vedro, mentr'io sospiro (Kreivi Almavivan aaria). - 5. Beethoven: Fidelio: I: Ha! welch ein Augenblick! (Pizarron aaria). - 6. Verdi: Un ballo in maschera: I: Alla vita che t'arride (Renaton aaria). - 7. Verdi: Un ballo in maschera: III: Alzati!...Eri tu (Renaton resit: & aaria). - //

F i s c h e r-D i e s k a u, Dietrich (bar); RSO Berlin (1-3); Berliner Philharmoniker (4); Bayerischeg Staatsorchester (5); Orch. der Deutschen Oper Berlin (6); L ö w l e i n, Hans (joht-1-3); L e i t n e r, Ferdinand (4); F r i c s a y, Ferenc (joht-5); P a t a n e, Giuseppe (joht-6,7); Bayerischer Staats-opernchor (5).

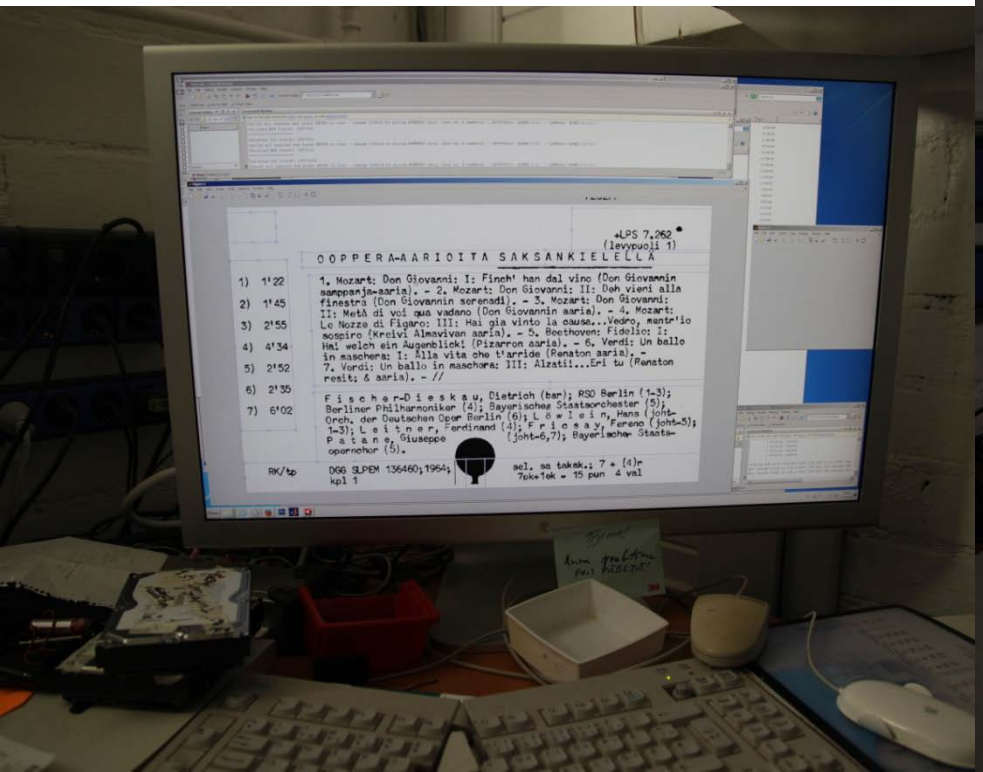
RK/tp

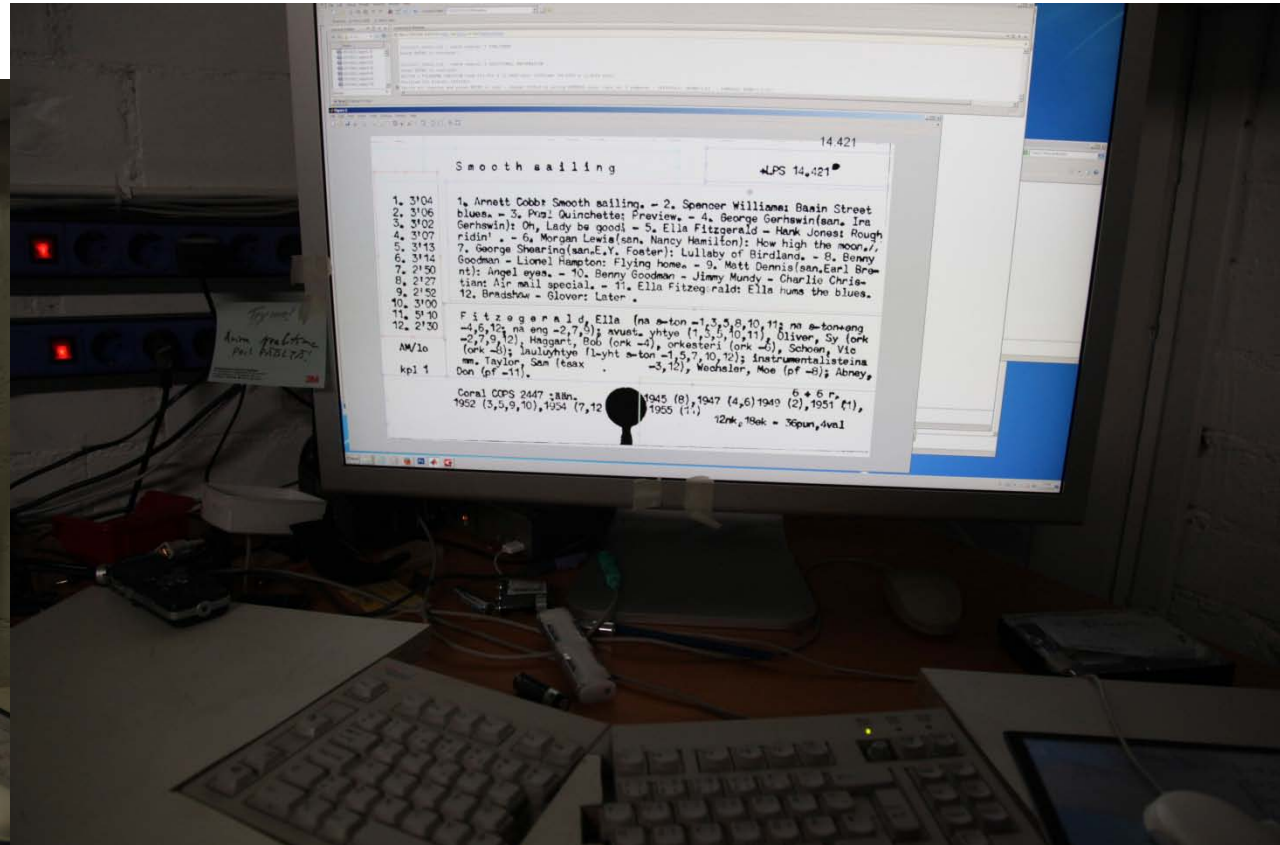
DGG SLPEM 136460;1964;
kpl 1

sel. sa takak.; 7 + (4)r
7pk+1ek - 15 pun 4 val

# 7. KAR Ltd.'s problem solving

- Too much between-card variation for 100 % automation
  - not possible to automate the region splitting process fully.
  - Numerous fine-tuned algorithms in the code -> runtime per card region splitting down to 15 – 30 s

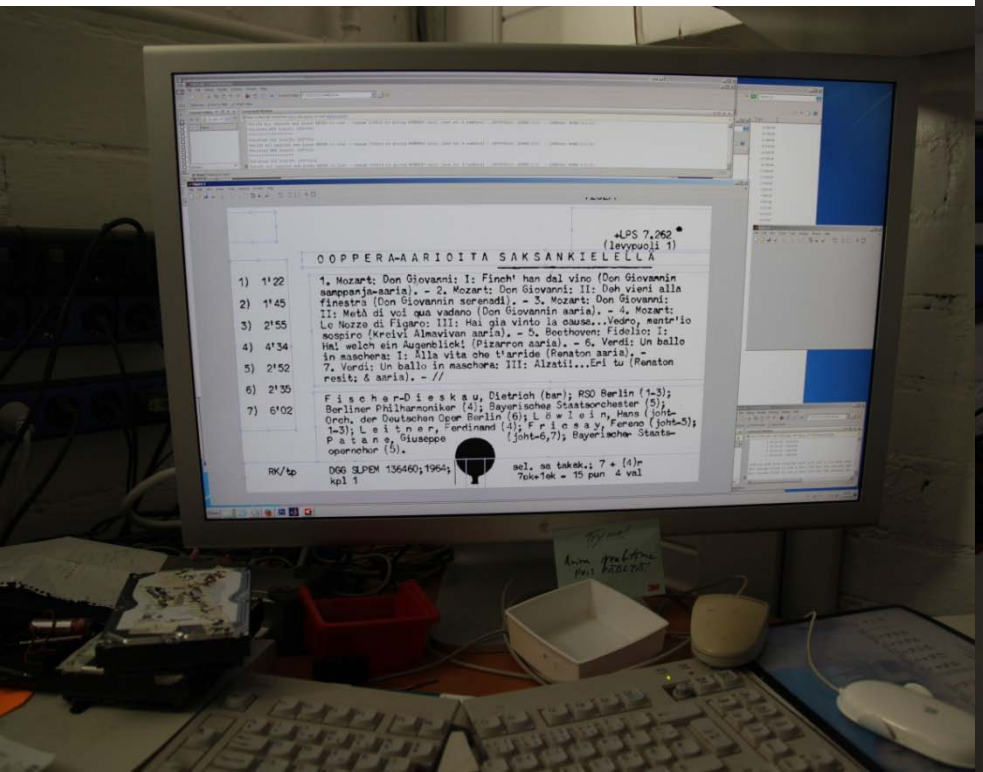


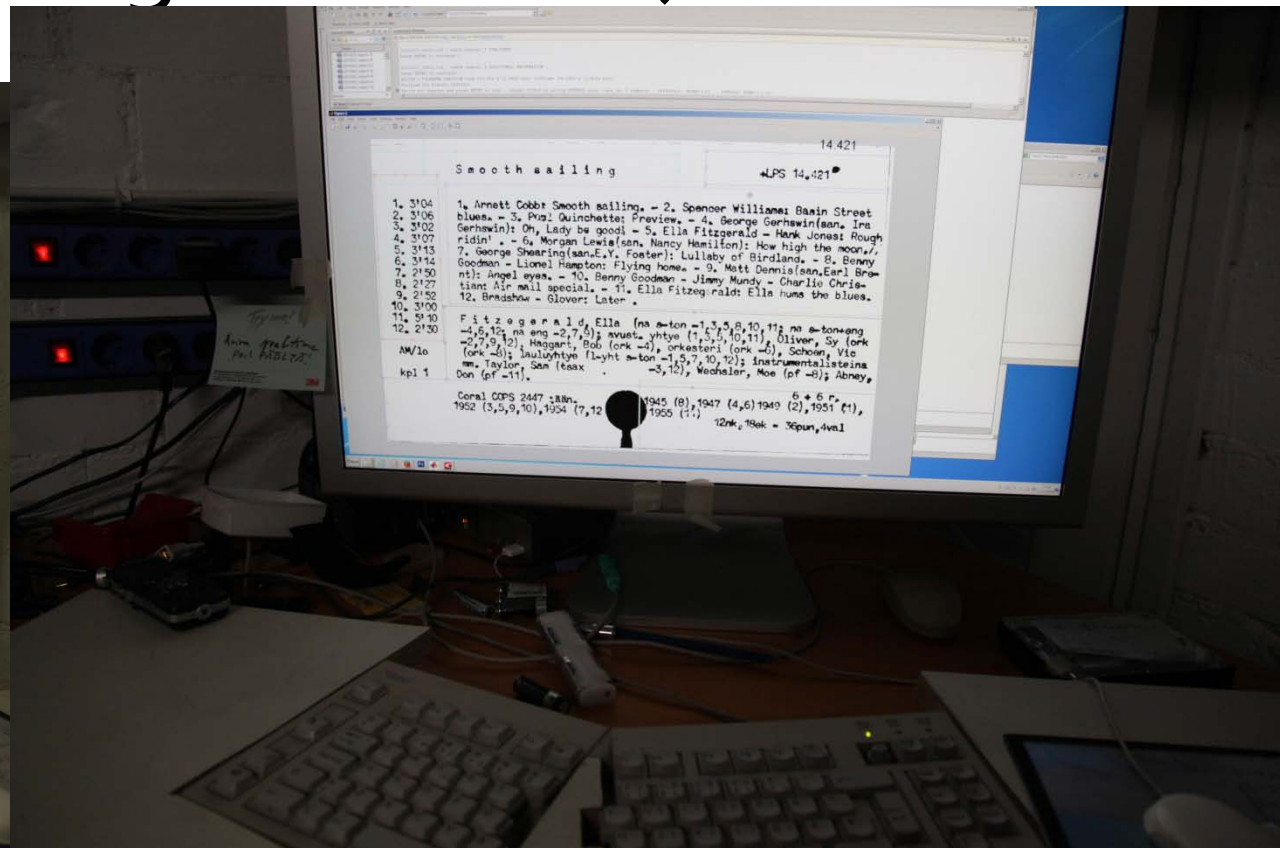


KAR

# 7. KAR Ltd.'s problem solving

- Real-time demo of the semi-automated region creation (= cropping the scanned image into 8 subimages for OCR)

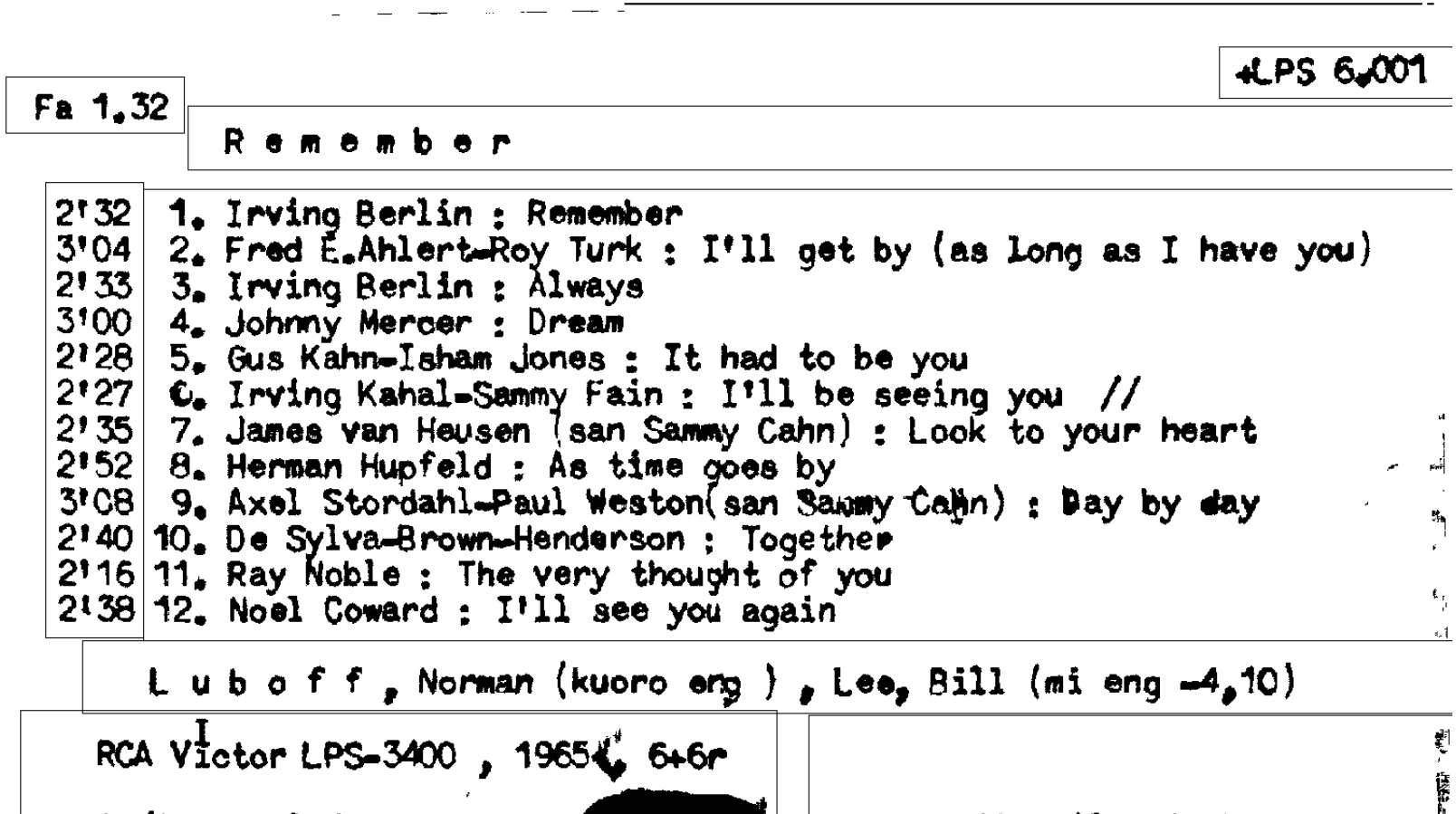




# 8. Improving the out-of-box OCR

KAR

- **Train the OCR** (= teach character patterns using the actual scanned documents) or not?



LPS 6.001

Fa 1.32

R e m e m b e r

| | |
|---|---|
| 2'32 | 1. Irving Berlin : Remember |
| 3'04 | 2. Fred E.Ahlert-Roy Turk : I'll get by (as long as I have you) |
| 2'33 | 3. Irving Berlin : Always |
| 3'00 | 4. Johnny Mercer : Dream |
| 2'28 | 5. Gus Kahn-Isham Jones : It had to be you |
| 2'27 | 6. Irving Kahal-Sammy Fain : I'll be seeing you // |
| 2'35 | 7. James van Heusen (san Sammy Cahn) : Look to your heart |
| 2'52 | 8. Herman Hupfeld : As time goes by |
| 3'08 | 9. Axel Stordahl-Paul Weston(san Sammy Cahn) : Day by day |
| 2'40 | 10. De Sylva-Brown-Henderson : Together |
| 2'16 | 11. Ray Noble : The very thought of you |
| 2'38 | 12. Noel Coward : I'll see you again |

L u b o f f , Norman (kuoro eng ) , Lee, Bill (mi eng -4,10)

RCA Victor LPS-3400 , 1965 6+6r

# 8. Tweak the OCR software!

KAR

- **Basic rules** set for the OCR software:
  - Only a **limited set of languages** allowed: English, Finnish, Frence, Germa, Latin, Spanish, Swedish
  - All languages include **the built-in dictionary and spell-check of the OCR** software
  - **Additional Finnish words** added to the dictionary.

# 8. Tweak the OCR software!

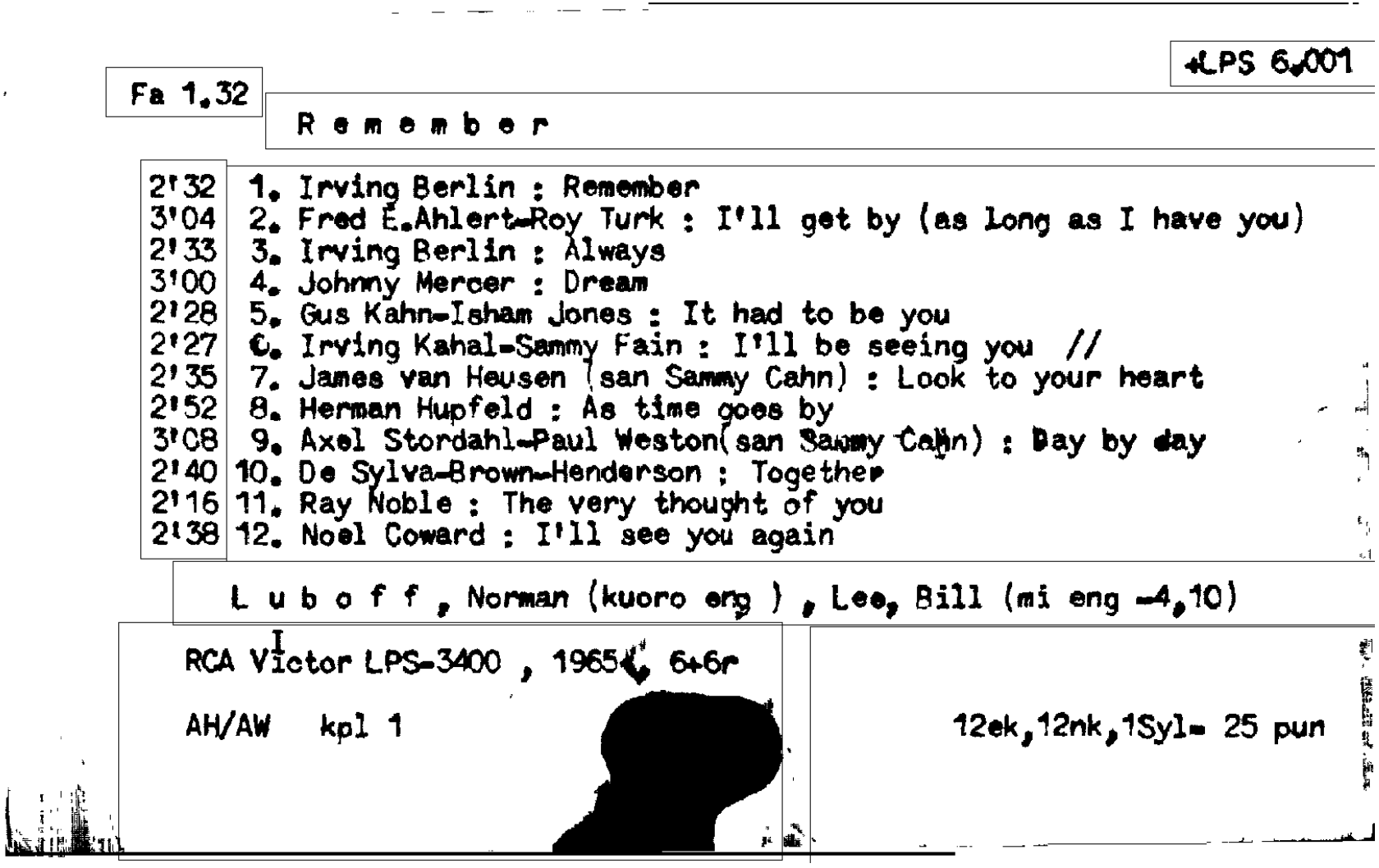
KAR

- Basic rules set for the OCR software (continued):
  - Only a **limited set of characters** (ca. 89 pieces): !"&'()+,./0123456789:;=?ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyzÄÜàáäåçèéöü
  - **Recognition area** = the complete area inside the region (= the scanned & cropped image)
  - **Output** as text files keeping line breaks, Unicode (UTF-8)

# 8. Training the OCR?

KAR

- Train the OCR or not?



+LPS 6,001

Fa 1,32

R e m e m b e r

| | |
|---|---|
| 2'32 | 1. Irving Berlin : Remember |
| 3'04 | 2. Fred E.Ahlert-Roy Turk : I'll get by (as Long as I have you) |
| 2'33 | 3. Irving Berlin : Always |
| 3'00 | 4. Johnny Mercer : Dream |
| 2'28 | 5. Gus Kahn-Isham Jones : It had to be you |
| 2'27 | 6. Irving Kahal-Sammy Fain : I'll be seeing you // |
| 2'35 | 7. James van Heusen (san Sammy Cahn) : Look to your heart |
| 2'52 | 8. Herman Hupfeld : As time goes by |
| 3'08 | 9. Axel Stordahl-Paul Weston(san Sammy Cahn) : Day by day |
| 2'40 | 10. De Sylva-Brown-Henderson : Together |
| 2'16 | 11. Ray Noble : The very thought of you |
| 2'38 | 12. Noel Coward : I'll see you again |

L u b o f f , Norman (kuoro eng ) , Lee, Bill (mi eng -4,10)

RCA Victor LPS-3400 , 1965 6+6r

AH/AW kpl 1

12ek,12nk,1Syl= 25 pun

# 8. Training the OCR?

KAR

- Out-of-the box results of the OCR

Fa 1.32

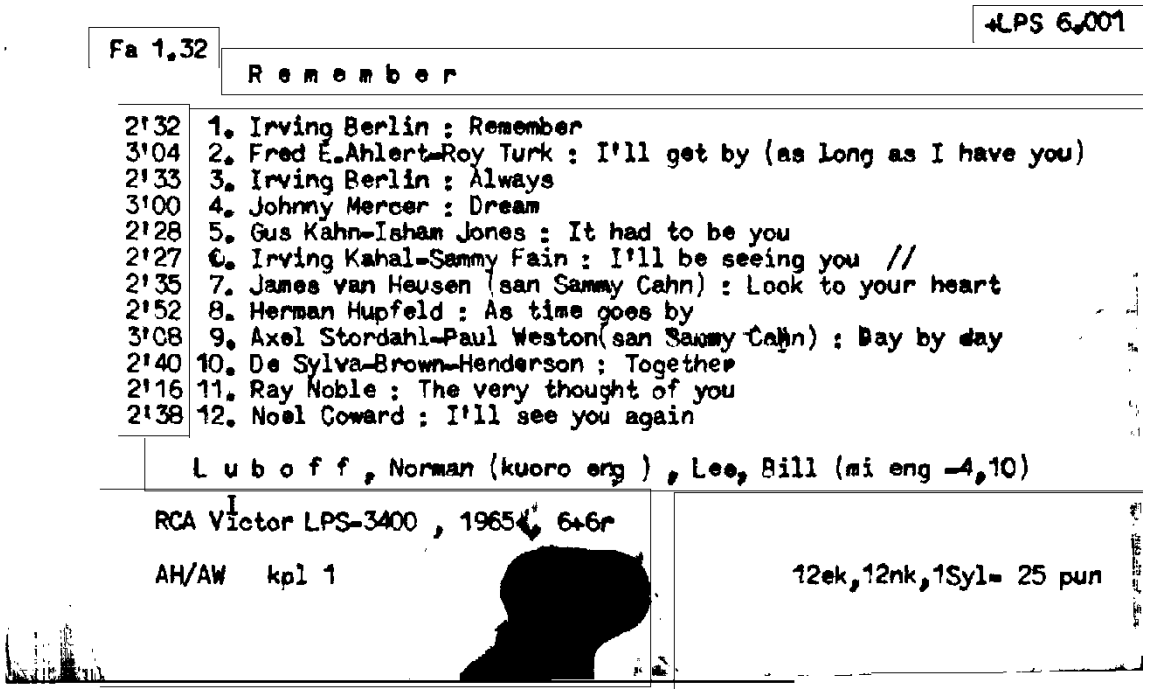
Remember

2r32 1,Irving Berlin ; Remember
3'04 2,Fred E.Ahlert-Roy Turk ; Ifll get by (as long as I have )
2'33 3,Irving Berlin : Always
3'00 4,Johnny Mercer ; Dream
2'28 5,Gus Kahn-Isham Jones : It had to be you
2'27 C,Irving Kahal-Sammy Fain : I'll beseeingyou//
2'35 7.James van Keusen (san Sarny Cahn): Looktoyour hea
2'52 8.Herman Hupfeld : As time goes by--
3fC8 9.Axel Stordahl-Paul Weston( san Sawwy -Ca n ) : 6ay
2'40 10,De Sylva-6rown-Henderson ; Together
2M6 11eRay Noble : The very thought of you
2138 12.Noel Coward : I111 see you again

t u b o f f , Norman (kuoro eng ) , Lee, Bill (mi eng -4,10)

RCA Victor LPS-3400 , 19654, 6t-6r
AH/AW
k^^-L

4.PS 6/Xtt



4LPS 6,001

Fa 1,32

R e m e m b e r

2'32 1. Irving Berlin : Remember
3'04 2. Fred E.Ahlert-Roy Turk : I'll get by (as long as I have you)
2'33 3. Irving Berlin : Always
3'00 4. Johnny Mercer : Dream
2'28 5. Gus Kahn-Isham Jones : It had to be you
2'27 6. Irving Kahal-Sammy Fain : I'll be seeing you //
2'35 7. James van Heusen (san Sammy Cahn) : Look to your heart
2'52 8. Herman Hupfeld : As time goes by
3'08 9. Axel Stordahl-Paul Weston(san Sammy Cahn) : Day by day
2'40 10. De Sylva-Brown-Henderson : Together
2'16 11. Ray Noble : The very thought of you
2'38 12. Noel Coward : I'll see you again

L u b o f f , Norman (kuoro eng ) , Lee, Bill (mi eng -4,10)

RCA Victor LPS-3400 , 1965[illegible] 6+6r

AH/AW kpl 1

12ek,12nk,1Syl- 25 pun

e
5
ft.
12ek,12nk,1Syl- 25 pun (
f
.-i--...1rii

# 8. Training the OCR

KAR

- Train the OCR or not (2)?

Ba 2
Cd 6
Cd 5.2
Cb 1

1) +LP 7.702
2) +LPS 6.099

V I V A L D I, Antonio

1) 10'30
2) 2'29
3'01
3'36
2'04+vr
11'18
3) 4'09
4) 7'30
5) 2'32
2'21
1'57+vr
6'53

1. Cessate, omai, kantaatti baritonille ja b c:lle. -
2. Sonaatti e-molli sellolle ja b c:lle (Largo, Allegro, Largo, Allegro). - //
3. La Senna festiggiante, "pastoraali-serenadi": II osa : Sinfonia (alkusoitto). -
4. Piango, gemo, sospiro, kantaatti baritonille ja b c:lle. -
5. Sonaatti c-molli viululle, sellolle ja b c:lle (Allegro, Largo, Allegro). -

M a l a g u t i, Laerte (bar- 1,4) ; R o v e d a, Egidio (vcl-2,4, 5) ; S c r o s o p p i, Antonio (vl- 5) ; S g r i z z i, Luciano (cem - 2,4,5) ; Società Cameristica di Lugano'n orkesteri (ork-1,3)
L o e h r e r, Edwin (joht- 1,3)

1) Nonesuch H 1088 ; saap, 1966 ; 5 + 5 r ; sel. eng ja tks it,eng takak.
2) Nonesuch H 71088

SL/SL

1) kpl 1,2 OK
2) kpl 1,2

5pk,8ek,5SyL = 12pun,13val

# 8. Training the OCR

KAR

- Out-of-the box results of the OCR(2)

f
/
ZBa 2
Cd 6
Cd 5.2
Cb 1
1)10'30
2)2'29
3'01
3'36
2'04+vr
3)4l09
4 7'30
' 5) 2'32
2'21
1'57+vr
I "TOJ
1

VIVALDI, Antonio

1. Cessate; omai. kantaatti baritonille ja b c:lie. -
2. Sonaatti e-moili sellolle ja b c:lie (Largo, Allegro, Larqo, Allegro). - //
3.La Senna festiggiante, "pastoraali-serenadi": II osa : Sinfonia (alkusoitto). -
4.Piango.gemo, sospiro, kantaatti baritonille ja b c:lle. -
5.Sonaatti c-molli viululle, sellolle ja b c:lle (Allegro, Largo, Allegro). -

Malaguti, Laerte (bar- 1,4) ; R o v e d a, Egidio (vol-2,4, 5);Scrosoppi. Antonio (vl- 5);Sgrizzi, Luciano (cem - 2,4,5) ; Societa Cameristica di Lugano'n orkesteri (o
L o e h r e r, Edwin (joht- 1,3)

3
!1) Nonesuch H 10S8 ; saap, 1966 ; 5 + 5 r ; sel. eng ja tks it,eng
2)Nonesuch H 71088takak.
1) kpl 1,2 OK5pk,8ek,5SvL-12pun,13val
SL/SL 2)kpli;i.g y-l
IJL

1)A.P 7.702
2)4L PS 6.099 v. i

Ba 2
Cd 6
Cd 5.2
Cb 1

1) 10'30
2) 2'29
3'01
3'36
2'04+vr
11'18
3) 4'09
4) 7'30
5) 2'32
2'21
1'57+vr
6'53

1) +LP 7.702
2) +LPS 6.099

V I V A L D I, Antonio

1. Cessate, omai, kantaatti baritonille ja b c:lle. -
2. Sonaatti e-molli sellolle ja b c:lle (Largo, Allegro, Largo, Allegro). - //
3. La Senna festiggiante, "pastoraali-serenadi": II osa : Sinfonia (alkusoitto). -
4.Piango, gemo, sospiro, kantaatti baritonille ja b c:lle. -
5. Sonaatti c-molli viululle, sellolle ja b c:lle (Allegro, Largo, Allegro). -

M a l a g u t i, Laerte (bar- 1,4) ; R o v e d a, Egidio (vcl-2,4, 5) ; S c r o s o p p i, Antonio (vl- 5) ; S g r i z z i, Luciano (cem - 2,4,5) ; Societa Cameristica di Lugano'n orkesteri (ork-1,
L o e h r e r, Edwin (joht- 1,3)

1) Nonesuch H 1088 ; saap, 1966 ; 5 + 5 r ; sel. eng ja tks it,eng
2) Nonesuch H 71088 takak.

SL/SL
1) kpl 1,2 OK
2) kpl 1,2

5pk,8ek,5SyL - 12pun,13val

5pk,8ek,5S[L] - 12pun,13val
-1

# 9. Training the OCR software with scanned YLE records' index cards

KAR

- All first 100 cards (LPS6001-LPS6099) learned per all (desired) characters
- Most of the cards learned several times
- The used typeface ("font") in the cards is very problematic
  - Letters "i" and "l" and number "1" are VERY SIMILAR to each other,
  - Letters "o" and "O" and number "O" are VERY SIMILAR to each other
  - There are many other similarities as well hindering the OCR process!

# 9. Training the OCR software with scanned YLE records' index cards

KAR

- Character templates (0-50) * 89 + **numerous ligatures**, eg.:

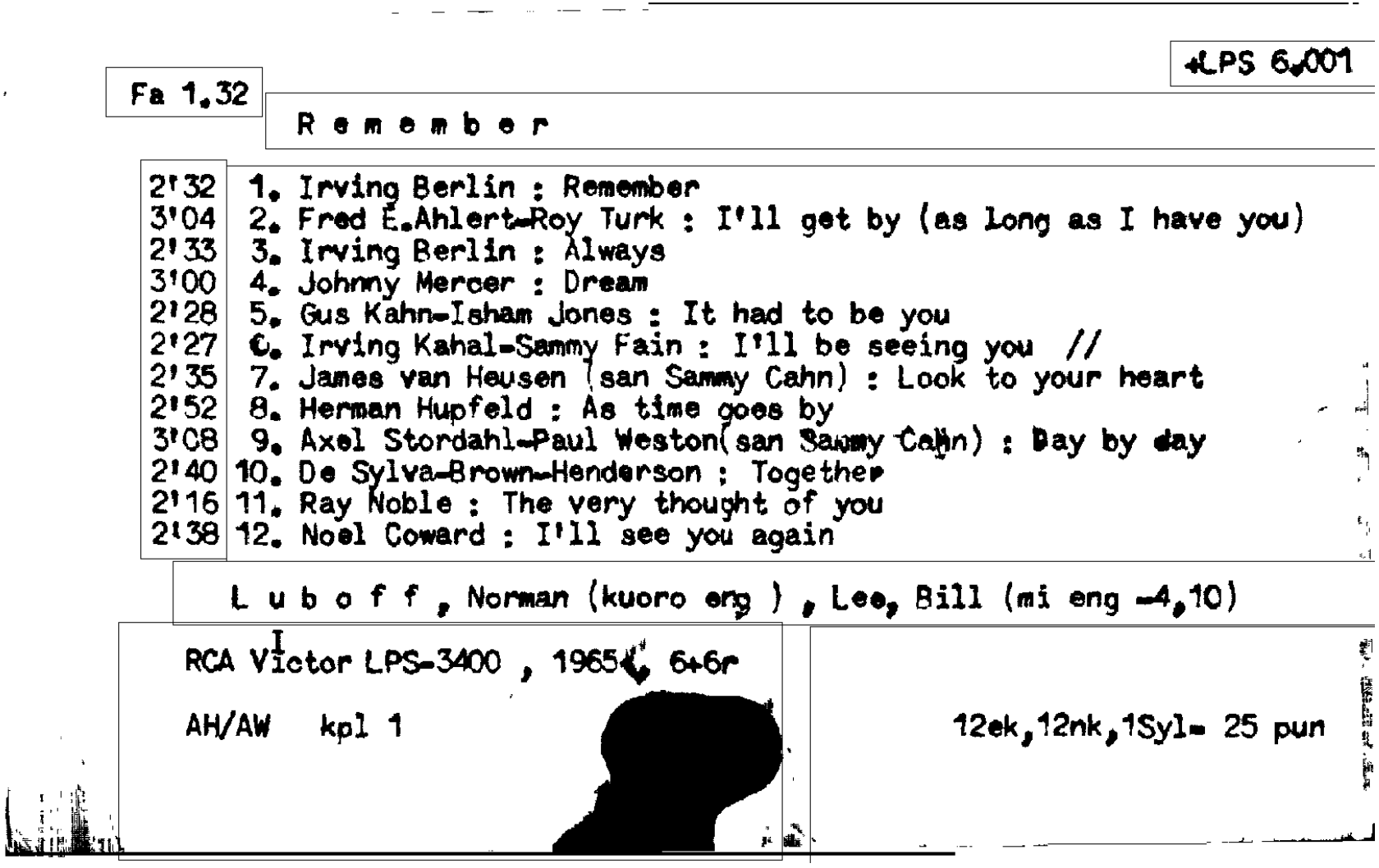
  11, 10, ll, lle, ie, ic, io, oc..

  [capitalized: 11, 10, LL, LLE, IE, IC, IO, OC ..]
- In total **ca. 4000 specially taught patterns** + built-in patterns used the OCR software.

# 9. Training the OCR software with scanned YLE records' index cards

KAR

- Scanned image (with regions)



+LPS 6.001

Fa 1.32

R e m e m b e r

| | |
|---|---|
| 2'32 | 1. Irving Berlin : Remember |
| 3'04 | 2. Fred E.Ahlert-Roy Turk : I'll get by (as Long as I have you) |
| 2'33 | 3. Irving Berlin : Always |
| 3'00 | 4. Johnny Mercer : Dream |
| 2'28 | 5. Gus Kahn-Isham Jones : It had to be you |
| 2'27 | 6. Irving Kahal-Sammy Fain : I'll be seeing you // |
| 2'35 | 7. James van Heusen (san Sammy Cahn) : Look to your heart |
| 2'52 | 8. Herman Hupfeld : As time goes by |
| 3'08 | 9. Axel Stordahl-Paul Weston(san Sammy Cahn) : Day by day |
| 2'40 | 10. De Sylva-Brown-Henderson : Together |
| 2'16 | 11. Ray Noble : The very thought of you |
| 2'38 | 12. Noel Coward : I'll see you again |

L u b o f f , Norman (kuoro eng ) , Lee, Bill (mi eng -4,10)

RCA Victor LPS-3400 , 1965 6+6r

AH/AW kpl 1

12ek,12nk,1Syl= 25 pun

# 9. Training the OCR software with scanned YLE records' index cards

KAR

- Untrained OCR's results

Fa 1.32

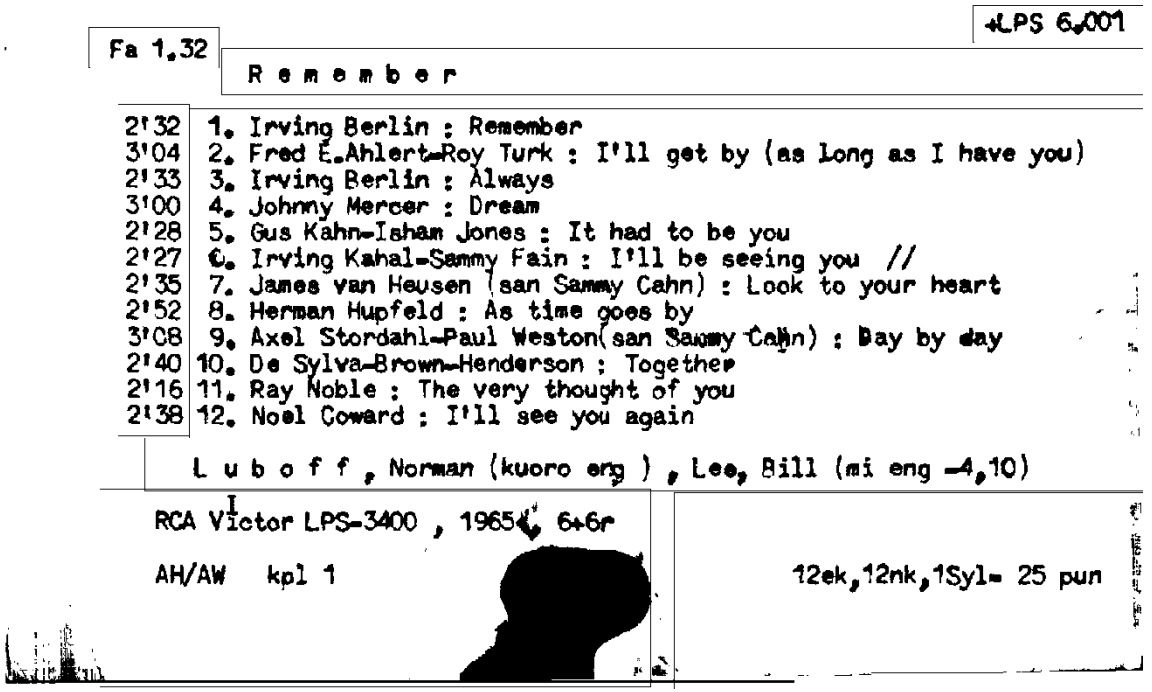
Remember

2r32 1,Irving Berlin ; Remember
3'04 2,Fred E.Ahlert-Roy Turk ; Ifll get by (as long as I have )
2'33 3,Irving Berlin : Always
3'00 4,Johnny Mercer ; Dream
2'28 5,Gus Kahn-Isham Jones : It had to be you
2'27 C,Irving Kahal-Sammy Fain : I'll beseeingyou//
2'35 7.James van Keusen (san Sarny Cahn): Looktoyour hea
2'52 8.Herman Hupfeld : As time goes by--
3fC8 9.Axel Stordahl-Paul Weston( san Sawwy -Ca n ) : 6ay
2'40 10,De Sylva-6rown-Henderson ; Together
2M6 11eRay Noble : The very thought of you
2138 12.Noel Coward : I111 see you again

t u b o f f , Norman (kuoro eng ) , Lee, Bill (mi eng -4,10)

RCA Victor LPS-3400 , 19654, 6t-6r
AH/AW
k^^-L

4.PS 6/Xtt



4LPS 6,001

Fa 1,32

R e m e m b e r

2'32 1. Irving Berlin : Remember
3'04 2. Fred E.Ahlert-Roy Turk : I'll get by (as long as I have you)
2'33 3. Irving Berlin : Always
3'00 4. Johnny Mercer : Dream
2'28 5. Gus Kahn-Isham Jones : It had to be you
2'27 6. Irving Kahal-Sammy Fain : I'll be seeing you //
2'35 7. James van Heusen (san Sammy Cahn) : Look to your heart
2'52 8. Herman Hupfeld : As time goes by
3'C8 9. Axel Stordahl-Paul Weston(san Sammy Cahn) : Day by day
2'40 10. De Sylva-Brown-Henderson : Together
2'16 11. Ray Noble : The very thought of you
2'38 12. Noel Coward : I'll see you again

L u b o f f , Norman (kuoro eng ) , Lee, Bill (mi eng -4,10)

RCA Victor LPS-3400 , 1965[illegible] 6+6r

AH/AW kpl 1

12ek,12nk,1Syl- 25 pun

e
5
ft.
12ek,12nk,1Syl- 25 pun (
f
.-i--...1rii
└.-J

# 9. Training the OCR software with scanned YLE records' index cards

KAR

- Fully trained OCR's results

Fa 1.32

Remember

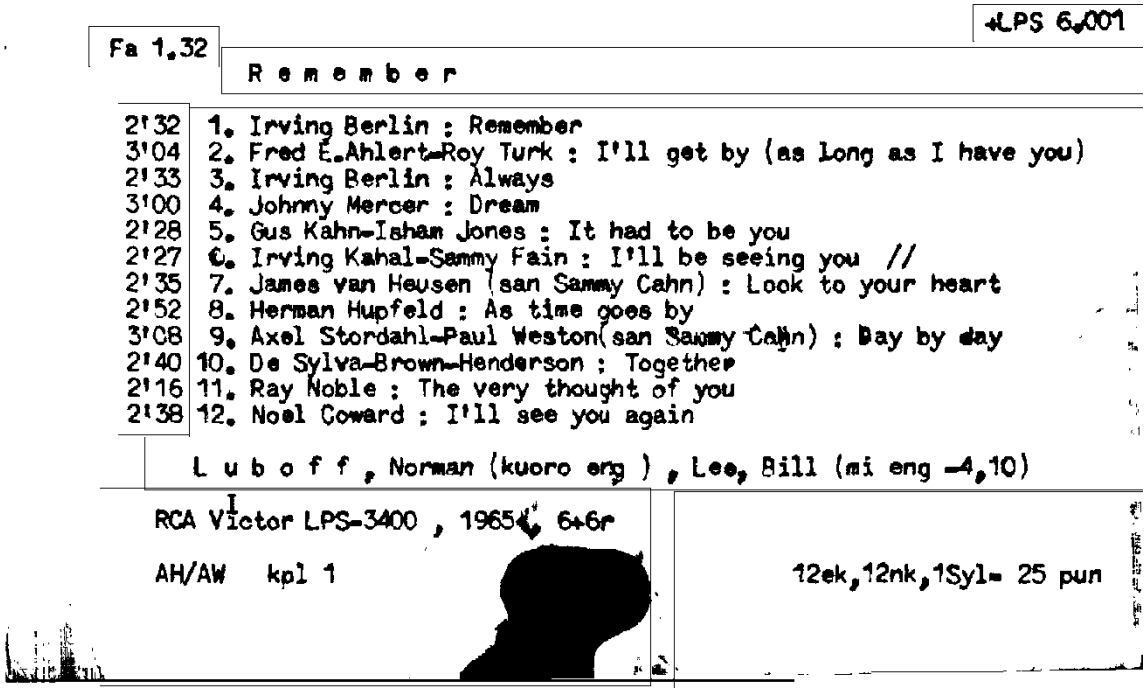
2'32 1.Irving Berlin : Remember
3'04 2.Fred E.Ählert-Roy Turk : I'll get by (as long as I have you)
2'33 3.Irving Berlin : Always
3'00 4.Johnny Mercer : Dream
2'28 5.Gus Kahn-Isham Jones : It had to be you
2'27 G. Irving Kahal-Sammy Fain : I'll beseeingyou//
2'35 7.James van Heusen (san Sammy Cahn):Looktoyour heart
2'52 8.Herman Hupfeld : As time goes by--
3'08 9.Axel Stordahl-Paul WestonCsan Sammy Cahn) s Day by
2'40 10.De Sylva-8rcwn-Henderson ; Together
2'16 11.Ray Noble : The very thought of you
2'38 12.Noel Coward : I'll see you again

L u b o f f , Norman (kuoro eng ) , Lee, Bill (mi eng -4,10)

RCA Victor LPS-3400 , 19654, 6+6r
AH/AW
L^^-1.

e
5
ft.
12ek,12nk,1Syl- 25 pun (
f
.-i--...1rii

+LPS 6.001



# 9. Training the OCR software with scanned YLE records' index cards

KAR

- Untrained vs. fully trained OCR's results

4.PS 6/Xtt

Fa 1.32

Remember

2r32 1,Irving Berlin ; Remember
3'04 2,Fred E.Ahlert-Roy Turk ; Ifll get by (as long as I have you)
2'33 3,Irving Berlin : Always
3'00 4,Johnny Mercer ; Dream
2'28 5,Gus Kahn-Isham Jones : It had to be you
2'27 C,Irving Kahal-Sammy Fain : I'll beseeingyou//
2'35 7.James van Keusen (san Sarny Cahn): Looktoyour heart
2'52 8.Herman Hupfeld : As time goes by--
3fC8 9.Axel Stordahl-Paul Weston( san Sawwy -Ca [n] ) : 6ay by day.
2'40 10,De Sylva-6rown-Henderson ; Together
2M6 11eRay Noble : The very thought of you
2138 12.Noel Coward : I111 see you again

t u b o f f , Norman (kuoro eng ) , Lee, Bill (mi eng -4,10)

RCA Victor LPS-3400 , 19654, 6t-6r
AH/AW
k^^-L

e
5
ft.
12ek,12nk,1Syl- 25 pun (
f
.-i--...1rii └.-J

+LPS 6.001

Fa 1.32

Remember

2'32 1.Irving Berlin : Remember
3'04 2.Fred E.Ählert-Roy Turk : I'll get by (as long as I have you)
2'33 3.Irving Berlin : Always
3'00 4.Johnny Mercer : Dream
2'28 5.Gus Kahn-Isham Jones : It had to be you
2'27 G. Irving Kahal-Sammy Fain : I'll beseeingyou//
2'35 7.James van Heusen (san Sammy Cahn):Looktoyour heart
2'52 8.Herman Hupfeld : As time goes by--
3'08 9.Axel Stordahl-Paul WestonCsan Sammy Cahn) s Day by day.
2'40 10.De Sylva-8rcwn-Henderson ; Together
2'16 11.Ray Noble : The very thought of you
2'38 12.Noel Coward : I'll see you again

L u b o f f , Norman (kuoro eng ) , Lee, Bill (mi eng -4,10)

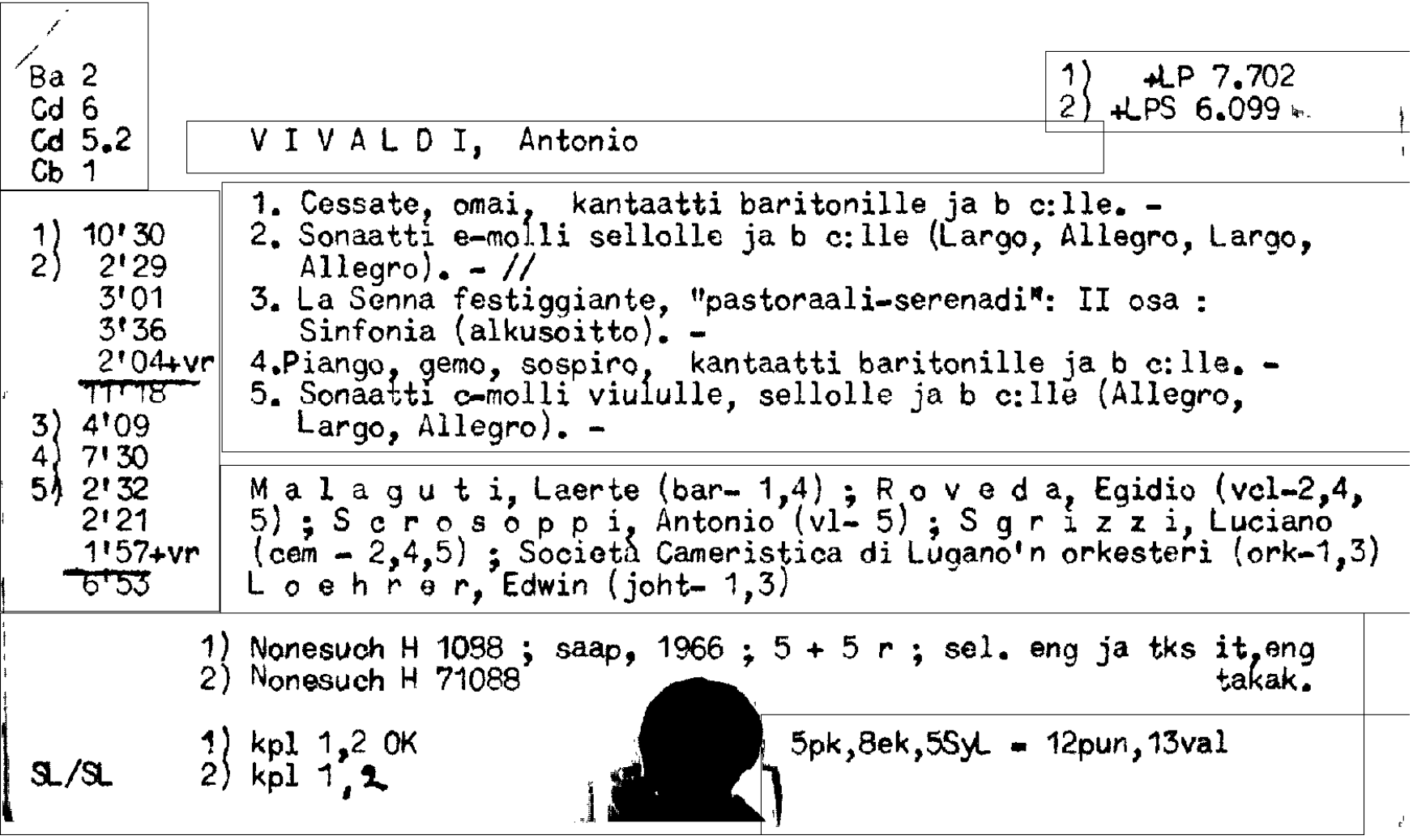
RCA Victor LPS-3400 , 19654, 6+6r
AH/AW
L^^-1.

e
5
ft.
12ek,12nk,1Syl- 25 pun (
f
.-i--...1rii └.-J

# 9. Training the OCR software with scanned YLE records' index cards

KAR

- Scanned image (with regions) (2)



Ba 2
Cd 6
Cd 5.2
Cb 1

1) +LP 7.702
2) +LPS 6.099

V I V A L D I, Antonio

1) 10'30
2) 2'29
3'01
3'36
2'04+vr
11'18
3) 4'09
4) 7'30
5) 2'32
2'21
1'57+vr
6'53

1. Cessate, omai, kantaatti baritonille ja b c:lle. -
2. Sonaatti e-molli sellolle ja b c:lle (Largo, Allegro, Largo, Allegro). - //
3. La Senna festiggiante, "pastoraali-serenadi": II osa : Sinfonia (alkusoitto). -
4. Piango, gemo, sospiro, kantaatti baritonille ja b c:lle. -
5. Sonaatti c-molli viululle, sellolle ja b c:lle (Allegro, Largo, Allegro). -

M a l a g u t i, Laerte (bar- 1,4) ; R o v e d a, Egidio (vcl-2,4, 5) ; S c r o s o p p i, Antonio (vl- 5) ; S g r i z z i, Luciano (cem - 2,4,5) ; Società Cameristica di Lugano'n orkesteri (ork-1,3)
L o e h r e r, Edwin (joht- 1,3)

1) Nonesuch H 1088 ; saap, 1966 ; 5 + 5 r ; sel. eng ja tks it,eng
2) Nonesuch H 71088 takak.

1) kpl 1,2 OK
SL/SL 2) kpl 1,2

5pk,8ek,5SyL = 12pun,13val

# 9. Training the OCR software with scanned YLE records' index cards

KAR

- Untrained OCR's results

f
/
ZBa 2
Cd 6
Cd 5.2
Cb 1
1)10'30
2)2'29
3'01
3'36
2'04+vr
3)4l09
4 7'30
' 5) 2'32
2'21
1'57+vr
I "TOJ
1

VIVALDI, Antonio

1. Cessate; omai. kantaatti baritonille ja b c:lie. -
2. Sonaatti e-moili sellolle ja b c:lie (Largo, Allegro, Larqo, Allegro). - //
3.La Senna festiggiante, "pastoraali-serenadi": II osa : Sinfonia (alkusoitto). -
4.Piango.gemo, sospiro, kantaatti baritonille ja b c:lle. -
5.Sonaatti c-molli viululle, sellolle ja b c:lle (Allegro, Largo, Allegro). -

Malaguti, Laerte (bar- 1,4) ; R o v e d a, Egidio (vol-2,4, 5);Scrosoppi. Antonio (vl- 5);Sgrizzi, Luciano (cem - 2,4,5) ; Societa Cameristica di Lugano'n orkesteri (o
L o e h r e r, Edwin (joht- 1,3)

3
!1) Nonesuch H 10S8 ; saap, 1966 ; 5 + 5 r ; sel. eng ja tks it,eng
2)Nonesuch H 71088takak.
1) kpl 1,2 OK5pk,8ek,5SvL-12pun,13val
SL/SL 2)kpli;i.g y-l
IJL

1)A.P 7.702
2)4L PS 6.099 v. i

Ba 2
Cd 6
Cd 5.2
Cb 1

1) 10'30
2) 2'29
3'01
3'36
2'04+vr
11'18
3) 4'09
4) 7'30
5) 2'32
2'21
1'57+vr
6'53

1) +LP 7.702
2) +LPS 6.099

V I V A L D I, Antonio

1. Cessate, omai, kantaatti baritonille ja b c:lle. -
2. Sonaatti e-molli sellolle ja b c:lle (Largo, Allegro, Largo, Allegro). - //
3. La Senna festiggiante, "pastoraali-serenadi": II osa : Sinfonia (alkusoitto). -
4.Piango, gemo, sospiro, kantaatti baritonille ja b c:lle. -
5. Sonaatti c-molli viululle, sellolle ja b c:lle (Allegro, Largo, Allegro). -

M a l a g u t i, Laerte (bar- 1,4) ; R o v e d a, Egidio (vcl-2,4,5) ; S c r o s o p p i, Antonio (vl- 5) ; S g r i z z i, Luciano (cem - 2,4,5) ; Societa Cameristica di Lugano'n orkesteri (ork-1,
L o e h r e r, Edwin (joht- 1,3)

1) Nonesuch H 1088 ; saap, 1966 ; 5 + 5 r ; sel. eng ja tks it,eng takak.
2) Nonesuch H 71088

SL/SL
1) kpl 1,2 OK
2) kpl 1,2

5pk,8ek,5SyL - 12pun,13val

5pk,8ek,5S[L] - 12pun,13val
-1

# 9. Training the OCR software with scanned YLE records' index cards

KAR

- Fully trained OCR's results

f
/
ZBa 2
Cd 6
Cd 5.2
Cb 1
1)10'30
2)2'29
3'01
3'36
2'04+vr
3)4'09
4 7' 30
' 5) 2'32
2'21
1'57+vr
l "ra
1

VIVALDI, Antonio

1. Cessate, omai, kantaatti baritonille ja b c:lie. -
2. Sonaatti e-rnolli sellolle ja b c:lie (Largo, Allegro, Largo, Allegro). - //
3.La Senna festiggiante, "pastoraali-serenadi": Il osa : Sinfonia (alkusoitto). -
4.Piango,gemo, sospiro, kantaatti baritonille ja b c:lie. -
5.Sonaatti c-molli viululle, sellolle ja b c:lie (Allegro, Largo, Allegro). -

Malaguti, Laerte (bar- 1,4) ; R o v e d a, Egidio (vcl-2,4, 5); S e r o s o p p i. Antonio (vl- 5);Sgrizzi, Luciano (cem - 2,4,5) ; Società Cameristica di Lugano'n orkesteri (o
L o e h r e r, Edwin (joht- 1,3)

3
!1) NonesuchH10S8 ;saap, 1966 ; 5 + 5 r ; sel. eng ja tksit,eng
2)NonesuchH71088takak.
1) kpl 1,2 OK5pk,8ek,5SyL-12pun,13val
SL/SL2)kpll'2
IJL

1)+LP 7.702
2)+LPS 6.099 v. i

Ba 2
Cd 6
Cd 5.2
Cb 1

1) 10'30
2) 2'29
3'01
3'36
2'04+vr
11'18
3) 4'09
4) 7'30
5) 2'32
2'21
1'57+vr
6'53

1) +LP 7.702
2) +LPS 6.099

V I V A L D I, Antonio

1. Cessate, omai, kantaatti baritonille ja b c:lle. -
2. Sonaatti e-molli sellolle ja b c:lle (Largo, Allegro, Largo, Allegro). - //
3. La Senna festiggiante, "pastoraali-serenadi": II osa : Sinfonia (alkusoitto). -
4.Piango, gemo, sospiro, kantaatti baritonille ja b c:lle. -
5. Sonaatti c-molli viululle, sellolle ja b c:lle (Allegro, Largo, Allegro). -

M a l a g u t i, Laerte (bar- 1,4) ; R o v e d a, Egidio (vcl-2, 5) ; S c r o s o p p i, Antonio (vl- 5) ; S g r i z z i, Luciano (cem - 2,4,5) ; Società Cameristica di Lugano'n orkesteri (ork-1, L o e h r e r, Edwin (joht- 1,3)

1) Nonesuch H 1088 ; saap, 1966 ; 5 + 5 r ; sel. eng ja tks it,eng
2) Nonesuch H 71088 takak.

SL/SL
1) kpl 1,2 OK
2) kpl 1,2

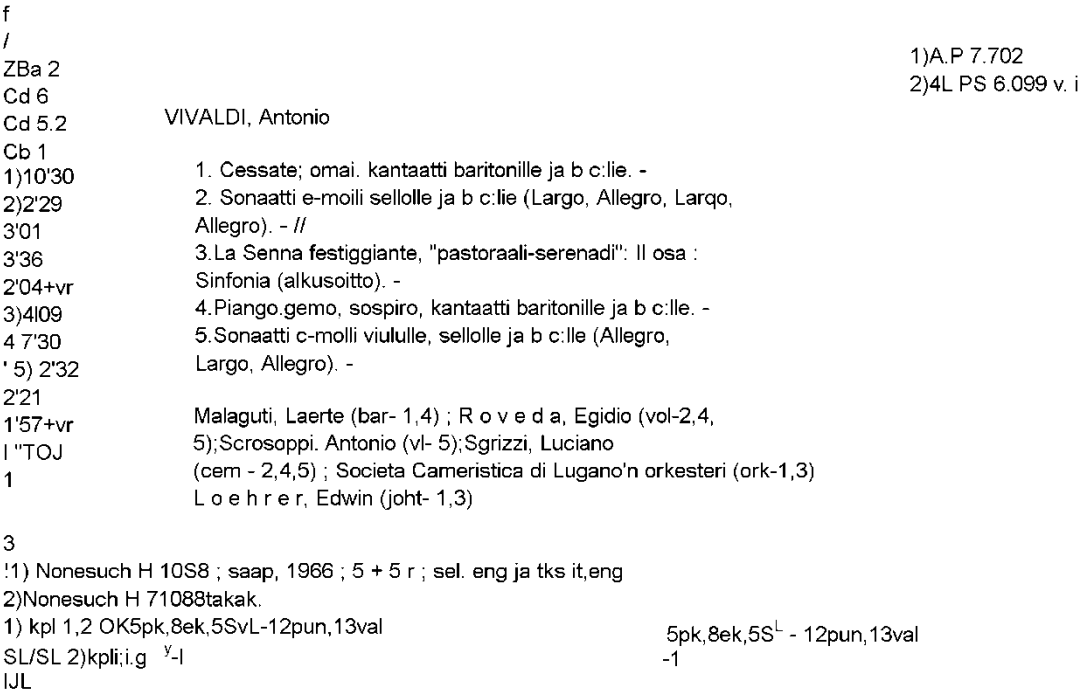
5pk,8ek,5SyL - 12pun,13val

5pk,8ek,5S[L] - 12pun,13val
-1

# 9. Training the OCR software with scanned YLE records' index cards

KAR

- Untrained vs. fully trained OCR's results



f
/
ZBa 2
Cd 6
Cd 5.2
Cb 1
1)10'30
2)2'29
3'01
3'36
2'04+vr
3)4l09
4 7'30
' 5) 2'32
2'21
1'57+vr
I "TOJ
1

1)A.P 7.702
2)4L PS 6.099 v. i

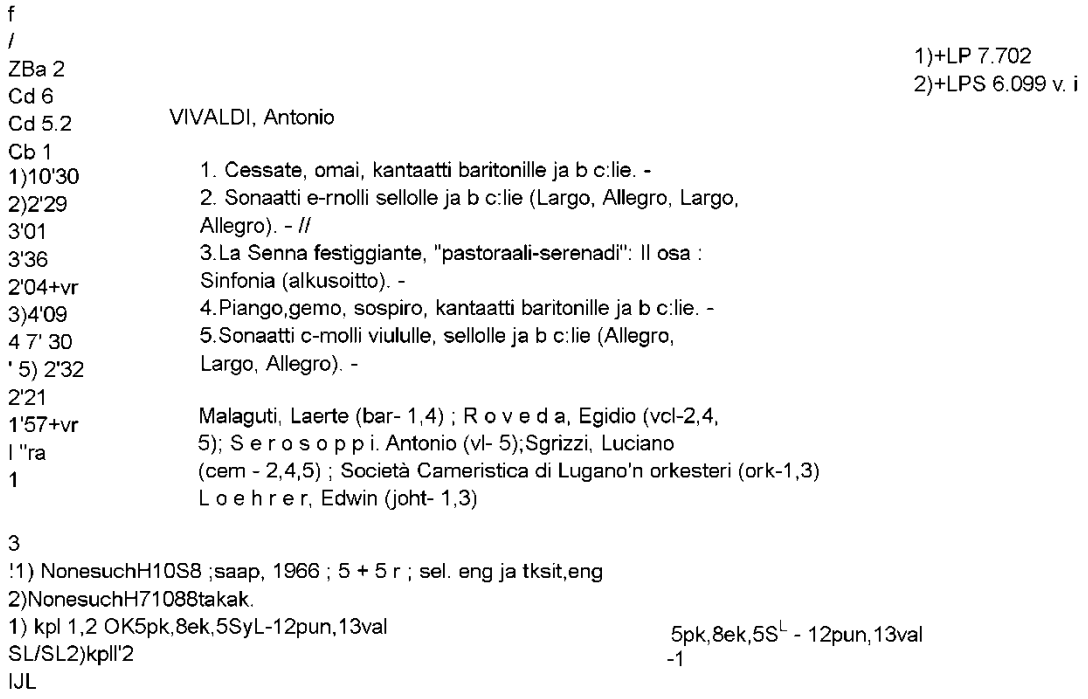
VIVALDI, Antonio

1. Cessate; omai. kantaatti baritonille ja b c:lie. -
2. Sonaatti e-moili sellolle ja b c:lie (Largo, Allegro, Larqo, Allegro). - //
3.La Senna festiggiante, "pastoraali-serenadi": Il osa : Sinfonia (alkusoitto). -
4.Piango.gemo, sospiro, kantaatti baritonille ja b c:lle. -
5.Sonaatti c-molli viululle, sellolle ja b c:lle (Allegro, Largo, Allegro). -

Malaguti, Laerte (bar- 1,4) ; R o v e d a, Egidio (vol-2,4, 5);Scrosoppi. Antonio (vl- 5);Sgrizzi, Luciano (cem - 2,4,5) ; Societa Cameristica di Lugano'n orkesteri (ork-1,3)
L o e h r e r, Edwin (joht- 1,3)

3
!1) Nonesuch H 10S8 ; saap, 1966 ; 5 + 5 r ; sel. eng ja tks it,eng
2)Nonesuch H 71088takak.
1) kpl 1,2 OK5pk,8ek,5SvL-12pun,13val
SL/SL 2)kpli;i.g $^{y}$-I
IJL

5pk,8ek,5S$^{L}$ - 12pun,13val
-1



f
/
ZBa 2
Cd 6
Cd 5.2
Cb 1
1)10'30
2)2'29
3'01
3'36
2'04+vr
3)4'09
4 7' 30
' 5) 2'32
2'21
1'57+vr
I "ra
1

1)+LP 7.702
2)+LPS 6.099 v. i

VIVALDI, Antonio

1. Cessate, omai, kantaatti baritonille ja b c:lie. -
2. Sonaatti e-rnolli sellolle ja b c:lie (Largo, Allegro, Largo, Allegro). - //
3.La Senna festiggiante, "pastoraali-serenadi": Il osa : Sinfonia (alkusoitto). -
4.Piango,gemo, sospiro, kantaatti baritonille ja b c:lie. -
5.Sonaatti c-molli viululle, sellolle ja b c:lie (Allegro, Largo, Allegro). -

Malaguti, Laerte (bar- 1,4) ; R o v e d a, Egidio (vcl-2,4, 5); S e r o s o p p i. Antonio (vl- 5);Sgrizzi, Luciano (cem - 2,4,5) ; Società Cameristica di Lugano'n orkesteri (ork-1,3)
L o e h r e r, Edwin (joht- 1,3)

3
!1) NonesuchH10S8 ;saap, 1966 ; 5 + 5 r ; sel. eng ja tksit,eng
2)NonesuchH71088takak.
1) kpl 1,2 OK5pk,8ek,5SyL-12pun,13val
SL/SL2)kpll'2
IJL

5pk,8ek,5S$^{L}$ - 12pun,13val
-1

Show more results

KAR

# 10. Making the OCR more effective

- The OCR software (the non-server versions, at least) has plenty of ~~limitations~~ features, e.g.
  - Too many pages (independent of actual size in pixels) cause it to jam (**memory extensive** application)
  - The software cannot use effectively multiple threads (**no multi-CPU support**)
  - The **software hangs** seldomly, but hangs alltogether
  - The software sometimes **cannot save its results** (OCRred files), causing a hang on the process
  - The software has **no shell support** for effective scripting (automation)

# 10. Making the OCR more effective

KAR

- The OCR software manufacturer has multiple versions of its software, including server based applications, but
  - with an additional price per OCRred page)

# 10. Making the OCR more effective

KAR

- To smooth the OCR process
  - Limit the page count to < 100 (per OCR document)
    - Running typically 9 YLE cards = 9*8 = 72 pages
  - Do nothing else with your computer when the OCR is running
  - Set a task for OCR software with all options & conditions preset
    - Do not allow preprocessing (when doing automated OCRing)

# 10. Automating the OCR software!

KAR

- To automate the OCR process
  - Create automated process, in which the OCR software always goes through the same folder with the same amount of images and have a master program to call it by window handling, mouse clicks and sort out the file handling
- **Show demo**

KAR

# 11. Outcome

- OCR software creates txt files that customer will parse further according to their needs
- For consistency checks, KAR Ltd. also parsed the data into Excel (XLS) spread sheet following the following structure:

| **Region information** | **1. region** | | | **2. region** | **3. region** |
|---|---|---|---|---|---|
| Region workname | DISCNRO (~ basename) | | | CLASS | SONGLEN |
| Description | stereo/mono | integer | side/ring | mixed codes | lenght of songs |
| 1. option | ***+LPS*** | ***12345*** | | ***Abc.123*** | ***1) XY'Zz*** |
| 2. option | ***+LP*** | | ***A*** | (several) | ***1. XY'Zz*** |
| following options | | | ***B1*** | | ***XY'Zz*** |
| | | | | | |
| *(Data can be missing)* | | | *(yes)* | *(yes)* | *(yes)* |

| **4. region** | **5. region** | **6. region** | **7. region** | **8. region** |
|---|---|---|---|---|
| TITLE | SONGNAME | ARTIST | PUBLISHER | ADDINFO |
| Disk title | song names | artist | publisher | additional info |
| ***Disk name*** | ***song name*** | ***Artist*** | ***Record co.*** | ***Writers ID*** |
| ***Composer*** | ***song name*** | ***composer*** | (several) | ***Julkaisukielet*** |
| (several) | (several) | (several) | | (several) |
| | | | | |
| *(yes)* | *(yes)* | *(yes)* | *(yes)* | *(yes)* |

This table is split after region 3 for layout reasons.

See results.

In **bold** and *italics* are typical examples of the OCRred text. Region 1 (DISCNRO) is split into 3 columns, others have only 1 column and may be empty.

# 12. Conclusions

KAR

- Digitizing work is straightforward in the game
  - technical problems can always be solved / made a compromise – practical dilemmas are worse
- OCR is a technical matter
  - one must accept (make compromise) that the quality of the originals sets the baseline;
  - 100 % correctness is nowhere ("should typos be corrected")

# 12. Conclusions

KAR

- How does one succeed in making successful optical character recognition (OCR) of scanned documents?
- How does one judge success?
  - Error rate to what?
  - At what cost?
- **A happy customer is the best measure of all – right? ☺**

# 13. Further work

KAR

- The work of KAR Ltd. does not stop here
  - More work *has been done already* on scanning and OCRing microfilms (fiches) applying the above methods adjusted to the new work / specifications

    see here

- KAR Ltd. has just moved to (5 x) bigger location (468 m2) with *more machines and playground*

  **WELCOME TO VISIT US AT ANYTIME!** ☺

Klaus A J Riederer, Dr. (tech),
KAR Ltd, CEO:

# You most likely have not learned a thing – but hopefully you have at least thought of something!